

02-16-00

Practitioner's Docket No. PERLIN-8

PATENT

10682 U.S. PTO
02/15/00

Preliminary Classification:

Proposed Class:

Subclass:

NOTE: "All applicants are requested to include a preliminary classification on newly filed patent applications. The preliminary classification, preferably class and subclass designations, should be identified in the upper right-hand corner of the letter of transmittal accompanying the application papers, for example 'Proposed Class 2, subclass 129.'" M.P.E.P. § 601, 7th ed.

10678 U.S. PTO
09/504389
02/15/00

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Box Patent Application
Assistant Commissioner for Patents
Washington, D.C. 20231

NEW APPLICATION TRANSMITTAL

Transmitted herewith for filing is the patent application of

Inventor(s): Mark W. Perlin

WARNING: 37 C.F.R. § 1.41(a)(1) points out:

"(a) A patent is applied for in the name or names of the actual inventor or inventors.

"(1) The inventorship of a nonprovisional application is that inventorship set forth in the oath or declaration as prescribed by § 1.63, except as provided for in § 1.53(d)(4) and § 1.63(d). If an oath or declaration as prescribed by § 1.63 is not filed during the pendency of a nonprovisional application, the inventorship is that inventorship set forth in the application papers filed pursuant to § 1.53(b), unless a petition under this paragraph accompanied by the fee set forth in § 1.17(l) is filed supplying or changing the name or names of the inventor or inventors."

For (title):
A METHOD AND SYSTEM FOR DNA ANALYSIS

CERTIFICATION UNDER 37 C.F.R. § 1.10*

(Express Mail label number is mandatory.)

(Express Mail certification is optional.)

I hereby certify that this New Application Transmittal and the documents referred to as attached therein are being deposited with the United States Postal Service on this date February 15, 2000, in an envelope as "Express Mail Post Office to Addressee," mailing Label Number EL396485536US, addressed to the: Assistant Commissioner for Patents, Washington, D.C. 20231.

Tracey L. Milka

(type or print name of person mailing paper)

Tracey L. Milka

Signature of person mailing paper

WARNING: Certificate of mailing (first class) or facsimile transmission procedures of 37 C.F.R. § 1.8 cannot be used to obtain a date of mailing or transmission for this correspondence.

***WARNING:** Each paper or fee filed by "Express Mail" **must** have the number of the "Express Mail" mailing label placed thereon prior to mailing. 37 C.F.R. § 1.10(b).

"Since the filing of correspondence under § 1.10 without the Express Mail mailing label thereon is an oversight that can be avoided by the exercise of reasonable care, requests for waiver of this requirement will **not** be granted on petition." Notice of Oct. 24, 1996, 60 Fed. Reg. 56,439, at 56,442.

1. Type of Application

This new application is for a(n)

(check one applicable item below)

- Original (nonprovisional)
- Design
- Plant

WARNING: Do not use this transmittal for a completion in the U.S. of an International Application under 35 U.S.C. § 371(c)(4), unless the International Application is being filed as a divisional, continuation or continuation-in-part application.

WARNING: Do not use this transmittal for the filing of a provisional application.

NOTE: If one of the following 3 items apply, then complete and attach ADDED PAGES FOR NEW APPLICATION TRANSMITTAL WHERE BENEFIT OF A PRIOR U.S. APPLICATION CLAIMED and a NOTIFICATION IN PARENT APPLICATION OF THE FILING OF THIS CONTINUATION APPLICATION.

- Divisional.
- Continuation.
- Continuation-in-part (C-I-P).

2. Benefit of Prior U.S. Application(s) (35 U.S.C. §§ 119(e), 120, or 121)

NOTE: A nonprovisional application may claim an invention disclosed in one or more prior filed copending nonprovisional applications or copending international applications designating the United States of America. In order for a nonprovisional application to claim the benefit of a prior filed copending nonprovisional application or copending international application designating the United States of America, each prior application must name as an inventor at least one inventor named in the later filed nonprovisional application and disclose the named inventor's invention claimed in at least one claim of the later filed nonprovisional application in the manner provided by the first paragraph of 35 U.S.C. § 112. Each prior application must also be:

- (i) An international application entitled to a filing date in accordance with PCT Article 11 and designating the United States of America; or
- (ii) Complete as set forth in § 1.51(b); or
- (iii) Entitled to a filing date as set forth in § 1.53(b) or § 1.53(d) and include the basic filing fee set forth in § 1.16; or
- (iv) Entitled to a filing date as set forth in § 1.53(b) and have paid therein the processing and retention fee set forth in § 1.21(l) within the time period set forth in § 1.53(f).

37 C.F.R. § 1.78(a)(1).

NOTE: If the new application being transmitted is a divisional, continuation or a continuation-in-part of a parent case, or where the parent case is an International Application which designated the U.S., or benefit of a prior provisional application is claimed, then check the following item and complete and attach ADDED PAGES FOR NEW APPLICATION TRANSMITTAL WHERE BENEFIT OF PRIOR U.S. APPLICATION(S) CLAIMED.

WARNING: If an application claims the benefit of the filing date of an earlier filed application under 35 U.S.C. §§ 120, 121 or 365(c), the 20-year term of that application will be based upon the filing date of the earliest U.S. application that the application makes reference to under 35 U.S.C. §§ 120, 121 or 365(c). (35 U.S.C. § 154(a)(2) does not take into account, for the determination of the patent term, any application on which priority is claimed under 35 U.S.C. §§ 119, 365(a) or 365(b).) For a c-i-p application, applicant should review whether any claim in the patent that will issue is supported by an earlier application and, if not, the applicant should consider canceling the reference to the earlier filed application. The term of a patent is not based on a claim-by-claim approach. See Notice of April 14, 1995, 60 Fed. Reg. 20,195, at 20,205.

WARNING: When the last day of pendency of a provisional application falls on a Saturday, Sunday, or Federal holiday within the District of Columbia, any nonprovisional application claiming benefit of the provisional application **must** be filed prior to the Saturday, Sunday, or Federal holiday within the District of Columbia. See 37 C.F.R. § 1.78(a)(3).

The new application being transmitted claims the benefit of prior U.S. application(s). Enclosed are ADDED PAGES FOR NEW APPLICATION TRANSMITTAL WHERE BENEFIT OF PRIOR U.S. APPLICATION(S) CLAIMED.

3. Papers Enclosed

A. Required for filing date under 37 C.F.R. § 1.53(b) (Regular) or 37 C.F.R. § 1.153 (Design) Application

81 Pages of specification
6 Pages of claims
14 Sheets of drawing

WARNING: DO NOT submit original drawings. A high quality copy of the drawings should be supplied when filing a patent application. The drawings that are submitted to the Office must be on strong, white, smooth, and non-shiny paper and meet the standards according to § 1.84. If corrections to the drawings are necessary, they should be made to the original drawing and a high-quality copy of the corrected original drawing then submitted to the Office. Only one copy is required or desired. For comments on proposed then-new 37 C.F.R. § 1.84, see Notice of March 9, 1988 (1990 O.G. 57-62).

NOTE: "Identifying indicia, if provided, should include the application number or the title of the invention, inventor's name, docket number (if any), and the name and telephone number of a person to call if the Office is unable to match the drawings to the proper application. This information should be placed on the back of each sheet of drawing a minimum distance of 1.5 cm. (5/8 inch) down from the top of the page . . ." 37 C.F.R. § 1.84(c).

(complete the following, if applicable)

The enclosed drawing(s) are photograph(s), and there is also attached a "PETITION TO ACCEPT PHOTOGRAPH(S) AS DRAWING(S)." 37 C.F.R. § 1.84(b).
 formal
 informal

B. Other Papers Enclosed

2 Pages of declaration and power of attorney
1 Pages of abstract
0 Other

4. Additional papers enclosed

Amendment to claims
 Cancel in this applications claims _____ before calculating the filing fee. (At least one original independent claim must be retained for filing purposes.)
 Add the claims shown on the attached amendment. (Claims added have been numbered consecutively following the highest numbered original claims.)
 Preliminary Amendment
 Information Disclosure Statement (37 C.F.R. § 1.98)
 Form PTO-1449 (PTO/SB/08A and 08B)
 Citations

- Declaration of Biological Deposit
- Submission of "Sequence Listing," computer readable copy and/or amendment pertaining thereto for biotechnology invention containing nucleotide and/or amino acid sequence.
- Authorization of Attorney(s) to Accept and Follow Instructions from Representative
- Special Comments
- Other

5. Declaration or oath (including power of attorney)

NOTE: A newly executed declaration is not required in a continuation or divisional application provided that the prior nonprovisional application contained a declaration as required, the application being filed is by all or fewer than all the inventors named in the prior application, there is no new matter in the application being filed, and a copy of the executed declaration filed in the prior application (showing the signature or an indication thereon that it was signed) is submitted. The copy must be accompanied by a statement requesting deletion of the names of person(s) who are not inventors of the application being filed. If the declaration in the prior application was filed under § 1.47, then a copy of that declaration must be filed accompanied by a copy of the decision granting § 1.47 status or, if a nonsigning person under § 1.47 has subsequently joined in a prior application, then a copy of the subsequently executed declaration must be filed. See 37 C.F.R. §§ 1.63(d)(1)–(3).

NOTE: A declaration filed to complete an application must be executed, identify the specification to which it is directed, identify each inventor by full name including family name and at least one given name, without abbreviation together with any other given name or initial, and the residence, post office address and country or citizenship of each inventor, and state whether the inventor is a sole or joint inventor. 37 C.F.R. § 1.63(a)(1)–(4).

NOTE: "The inventorship of a nonprovisional application is that inventorship set forth in the oath or declaration as prescribed by § 1.62, except as provided for in § 1.53(d)(4) and § 1.63(d). If an oath or declaration as prescribed by § 1.63 is not filed during the pendency of a nonprovisional application, the inventorship is that inventorship set forth in the application papers filed pursuant to § 1.53(b), unless a petition under this paragraph accompanied by the fee set forth in § 1.17(f) is filed supplying or changing the name or names of the inventor or inventors." 37 C.F.R. § 1.41(a)(1).

Enclosed

Executed by

(check all applicable boxes)

- inventor(s).
- legal representative of inventor(s).
37 C.F.R. §§ 1.42 or 1.43.
- joint inventor or person showing a proprietary interest on behalf of inventor who refused to sign or cannot be reached.
 - This is the petition required by 37 C.F.R. § 1.47 and the statement required by 37 C.F.R. § 1.47 is also attached. See item 13 below for fee.
- Not Enclosed.

NOTE: Where the filing is a completion in the U.S. of an International Application or where the completion of the U.S. application contains subject matter in addition to the International Application, the application may be treated as a continuation or continuation-in-part, as the case may be, utilizing ADDED PAGE FOR NEW APPLICATION TRANSMITTAL WHERE BENEFIT OF PRIOR U.S. APPLICATION CLAIMED.

- Application is made by a person authorized under 37 C.F.R. § 1.41(c) on behalf of all the above named inventor(s).

(The declaration or oath, along with the surcharge required by 37 C.F.R. § 1.16(e) can be filed subsequently).

Showing that the filing is authorized.
(not required unless called into question. 37 C.F.R. § 1.41(d))

6. Inventorship Statement

WARNING: If the named inventors are each not the inventors of all the claims an explanation, including the ownership of the various claims at the time the last claimed invention was made, should be submitted.

The inventorship for all the claims in this application are:

The same.

or

Not the same. An explanation, including the ownership of the various claims at the time the last claimed invention was made,
 is submitted.
 will be submitted.

7. Language

NOTE: An application including a signed oath or declaration may be filed in a language other than English. An English translation of the non-English language application and the processing fee of \$130.00 required by 37 C.F.R. § 1.17(k) is required to be filed with the application, or within such time as may be set by the Office. 37 C.F.R. § 1.52(d).

English

Non-English

The attached translation includes a statement that the translation is accurate. 37 C.F.R. § 1.52(d).

8. Assignment

An assignment of the invention to _____

is attached. A separate "COVER SHEET FOR ASSIGNMENT (DOCUMENT) ACCOMPANYING NEW PATENT APPLICATION" or FORM PTO 1595 is also attached.
 will follow.

NOTE: "If an assignment is submitted with a new application, send two separate letters-one for the application and one for the assignment." Notice of May 4, 1990 (1114 O.G. 77-78).

WARNING: A newly executed "CERTIFICATE UNDER 37 C.F.R. § 3.73(b)" must be filed when a continuation-in-part application is filed by an assignee. Notice of April 30, 1993, 1150 O.G. 62-64.

(New Application Transmittal [4-1]—page 5 of 11)

9. Certified Copy

Certified copy(ies) of application(s)

| Country | Appn. No. | Filed |
|---------|-----------|-------|
| Country | Appn. No. | Filed |
| Country | Appn. No. | Filed |

from which priority is claimed

is (are) attached.
 will follow.

NOTE: The foreign application forming the basis for the claim for priority must be referred to in the oath or declaration. 37 C.F.R. § 1.55(a) and 1.63.

NOTE: This item is for any foreign priority for which the application being filed directly relates. If any parent U.S. application or International Application from which this application claims benefit under 35 U.S.C. § 120 is itself entitled to priority from a prior foreign application, then complete item 18 on the ADDED PAGES FOR NEW APPLICATION TRANSMITTAL WHERE BENEFIT OF PRIOR U.S. APPLICATION(S) CLAIMED.

10. Fee Calculation (37 C.F.R. § 1.16)**A. Regular application**

| CLAIMS AS FILED | | | | |
|--|--------------|------|----------------------------------|--------|
| Number filed | Number Extra | Rate | Basic Fee 37 C.F.R. § 1.16(a) | |
| | | | \$760.00 | |
| Total | | | | |
| Claims (37 C.F.R. § 1.16(c)) | 22 - 20 = 2 | × | \$ 18.00 | 36.00 |
| Independent Claims (37 C.F.R. § 1.16(b)) | 6 - 3 = 3 | × | \$ 78.00 | 234.00 |
| Multiple dependent claim(s), if any (37 C.F.R. § 1.16(d)) | | + | \$260.00 | |

Amendment cancelling extra claims is enclosed.
 Amendment deleting multiple-dependencies is enclosed.
 Fee for extra claims is not being paid at this time.

NOTE: If the fees for extra claims are not paid on filing they must be paid or the claims cancelled by amendment, prior to the expiration of the time period set for response by the Patent and Trademark Office in any notice of fee deficiency. 37 C.F.R. § 1.16(d).

Filing Fee Calculation \$ 960.00

**B. Design application
(\$310.00—37 C.F.R. § 1.16(f))**

Filing Fee Calculation \$

C. Plant application
(\$480.00—37 C.F.R. § 1.16(g))

Filing fee calculation \$ _____

11. Small Entity Statement(s)

Statement(s) that this is a filing by a small entity under 37 C.F.R. § 1.9 and 1.27 is (are) attached.

WARNING: "Status as a small entity must be specifically established in each application or patent in which the status is available and desired. Status as a small entity in one application or patent does not affect any other application or patent, including applications or patents which are directly or indirectly dependent upon the application or patent in which the status has been established. The refiling of an application under § 1.53 as a continuation, division, or continuation-in-part (including a continued prosecution application under § 1.53(d)), or the filing of a reissue application requires a new determination as to continued entitlement to small entity status for the continuing or reissue application. A nonprovisional application claiming benefit under 35 U.S.C. § 119(e), 120, 121, or 365(c) of a prior application, or a reissue application may rely on a statement filed in the prior application or in the patent if the nonprovisional application or the reissue application includes a reference to the statement in the prior application or in the patent or includes a copy of the statement in the prior application or in the patent and status as a small entity is still proper and desired. The payment of the small entity basic statutory filing fee will be treated as such a reference for purposes of this section." 37 C.F.R. § 1.28(a)(2).

WARNING: "Small entity status must not be established when the person or persons signing the . . . statement can unequivocally make the required self-certification." M.P.E.P., § 509.03, 6th ed., rev. 2, July 1996 (emphasis added).

(complete the following, if applicable)

Status as a small entity was claimed in prior application

_____ / _____, filed on _____, from which benefit is being claimed for this application under:

35 U.S.C. § 119(e),
 120,
 121,
 365(c),

and which status as a small entity is still proper and desired.

A copy of the statement in the prior application is included.

Filing Fee Calculation (50% of A, B or C above)

\$ 480.00

NOTE: Any excess of the full fee paid will be refunded if small entity status is established and a refund request are filed within 2 months of the date of timely payment of a full fee. The two-month period is not extendable under § 1.136. 37 C.F.R. § 1.28(a).

12. Request for International-Type Search (37 C.F.R. § 1.104(d))

(complete, if applicable)

Please prepare an international-type search report for this application at the time when national examination on the merits takes place.

13. Fee Payment Being Made at This Time

Not Enclosed

No filing fee is to be paid at this time.
(*This and the surcharge required by 37 C.F.R. § 1.16(e) can be paid subsequently.*)

Enclosed

Filing fee \$ 480.00

Recording assignment
(\$40.00; 37 C.F.R. § 1.21(h))
(See attached "COVER SHEET FOR
ASSIGNMENT ACCOMPANYING NEW
APPLICATION".) \$ _____

Petition fee for filing by other than all the
inventors or person on behalf of the inventor
where inventor refused to sign or cannot be
reached
(\$130.00; 37 C.F.R. §§ 1.47 and 1.17(l)) \$ _____

For processing an application with a
specification in
a non-English language
(\$130.00; 37 C.F.R. §§ 1.52(d) and 1.17(k)) \$ _____

Processing and retention fee
(\$130.00; 37 C.F.R. §§ 1.53(d) and 1.21(l)) \$ _____

Fee for international-type search report
(\$40.00; 37 C.F.R. § 1.21(e)) \$ _____

NOTE: 37 C.F.R. § 1.21(l) establishes a fee for processing and retaining any application that is abandoned for failing to complete the application pursuant to 37 C.F.R. § 1.53(f) and this, as well as the changes to 37 C.F.R. §§ 1.53 and 1.78(a)(1), indicate that in order to obtain the benefit of a prior U.S. application, either the basic filing fee must be paid, or the processing and retention fee of § 1.21(l) must be paid, within 1 year from notification under § 53(f).

Total fees enclosed \$ 480.00

14. Method of Payment of Fees

Check in the amount of \$ 480.00

Charge Account No. _____ in the amount of
\$ _____

A duplicate of this transmittal is attached.

NOTE: Fees should be itemized in such a manner that it is clear for which purpose the fees are paid. 37 C.F.R. § 1.22(b).

15. Authorization to Charge Additional Fees

WARNING: If no fees are to be paid on filing, the following items should not be completed.

WARNING: Accurately count claims, especially multiple dependent claims, to avoid unexpected high charges, if extra claim charges are authorized.

The Commissioner is hereby authorized to charge the following additional fees by this paper and during the entire pendency of this application to Account No. 19-0737:

37 C.F.R. § 1.16(a), (f) or (g) (filing fees)

37 C.F.R. § 1.16(b), (c) and (d) (presentation of extra claims)

NOTE: Because additional fees for excess or multiple dependent claims not paid on filing or on later presentation must only be paid or these claims cancelled by amendment prior to the expiration of the time period set for response by the PTO in any notice of fee deficiency (37 C.F.R. § 1.16(d)), it might be best not to authorize the PTO to charge additional claim fees, except possibly when dealing with amendments after final action.

37 C.F.R. § 1.16(e) (surcharge for filing the basic filing fee and/or declaration on a date later than the filing date of the application)

37 C.F.R. § 1.17(a)(1)–(5) (extension fees pursuant to § 1.136(a)).

37 C.F.R. § 1.17 (application processing fees)

NOTE: ". . . A written request may be submitted in an application that is an authorization to treat any concurrent or future reply, requiring a petition for an extension of time under this paragraph for its timely submission, as incorporating a petition for extension of time for the appropriate length of time. An authorization to charge all required fees, fees under § 1.17, or all required extension of time fees will be treated as a constructive petition for an extension of time in any concurrent or future reply requiring a petition for an extension of time under this paragraph for its timely submission. Submission of the fee set forth in § 1.17(a) will also be treated as a constructive petition for an extension of time in any concurrent reply requiring a petition for an extension of time under this paragraph for its timely submission." 37 C.F.R. § 1.136(a)(3).

37 C.F.R. § 1.18 (issue fee at or before mailing of Notice of Allowance, pursuant to 37 C.F.R. § 1.311(b))

NOTE: Where an authorization to charge the issue fee to a deposit account has been filed before the mailing of a Notice of Allowance, the issue fee will be automatically charged to the deposit account at the time of mailing the notice of allowance. 37 C.F.R. § 1.311(b).

NOTE: 37 C.F.R. § 1.28(b) requires "Notification of any change in status resulting in loss of entitlement to small entity status must be filed in the application . . . prior to paying, or at the time of paying, . . . the issue fee. . ." From the wording of 37 C.F.R. § 1.28(b), (a) notification of change of status must be made even if the fee is paid as "other than a small entity" and (b) no notification is required if the change is to another small entity.

16. Instructions as to Overpayment

NOTE: ". . . Amounts of twenty-five dollars or less will not be returned unless specifically requested within a reasonable time, nor will the payer be notified of such amounts; amounts over twenty-five dollars may be returned by check or, if requested, by credit to a deposit account." 37 C.F.R. § 1.26(a).

Credit Account No. 19-0737
 Refund

0954389-031500



SIGNATURE OF PRACTITIONER

Ansel M. Schwartz

(type or print name of attorney)

One Sterling Plaza
201 N. Craig Street

P.O. Address

Suite 304
Pittsburgh, PA 15213

Reg. No. 30,587

Tel. No. (412) 621-9222

Customer No.

Incorporation by reference of added pages

(check the following item if the application in this transmittal claims the benefit of prior U.S. application(s) (including an international application entering the U.S. stage as a continuation, divisional or C-I-P application) and complete and attach the ADDED PAGES FOR NEW APPLICATION TRANSMITTAL WHERE BENEFIT OF PRIOR U.S. APPLICATION(S) CLAIMED)

Plus Added Pages for New Application Transmittal Where Benefit of Prior U.S. Application(s) Claimed

Number of pages added _____

Plus Added Pages for Papers Referred to in Item 4 Above

Number of pages added _____

Plus added pages deleting names of inventor(s) named in prior application(s) who is/are no longer inventor(s) of the subject matter claimed in this application.

Number of pages added _____

Plus "Assignment Cover Letter Accompanying New Application"

Number of pages added _____

Statement Where No Further Pages Added

(if no further pages form a part of this Transmittal, then end this Transmittal with this page and check the following item)

This transmittal ends with this page.

Attorney's Docket No. PERLIN-8

PATENT

Applicant or Patentee: Mark W. Perlin

Application or Patent No.: /

Filed or Issued: _____

For: A METHOD AND SYSTEM FOR DNA ANALYSIS

**VERIFIED STATEMENT (DECLARATION) CLAIMING SMALL ENTITY
STATUS (37 CFR 1.9(f) and 1.27(b))—INDEPENDENT INVENTOR**

As a below named inventor, I hereby declare that I qualify as an independent inventor, as defined in 37 CFR 1.9(c), for purposes of paying reduced fees under Sections 41(a) and (b) of Title 35, United States Code, to the Patent and Trademark Office with regard to the invention entitled A METHOD AND SYSTEM FOR DNA ANALYSIS

described in

the specification filed herewith.
 application no. / _____, filed _____.
 patent no. _____, issued _____.

I have not assigned, granted, conveyed or licensed, and am under no obligation under contract or law to assign, grant, convey or license, any rights in the invention to any person who could not be classified as an independent inventor under 37 CFR 1.9(c), if that person had made the invention, or to any concern that would not qualify as a small business concern under 37 CFR 1.9(d), or a nonprofit organization under 37 CFR 1.9(e).

Each person, concern or organization to which I have assigned, granted, conveyed, or licensed or am under an obligation under contract or law to assign, grant, convey, or license any rights in the invention is listed below:

no such person, concern, or organization.
 persons, concerns or organizations listed below *

**NOTE: Separate verified statements are required from each named person, concern or organization having rights to the invention affering to their status as small entities. (37 CFR 1.27)*

FULL NAME _____

ADDRESS _____

INDIVIDUAL SMALL BUSINESS CONCERN NONPROFIT ORGANIZATION

FULL NAME _____

ADDRESS _____

INDIVIDUAL SMALL BUSINESS CONCERN NONPROFIT ORGANIZATION

FULL NAME _____

ADDRESS _____

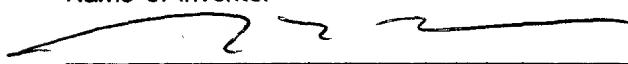
INDIVIDUAL SMALL BUSINESS CONCERN NONPROFIT ORGANIZATION

I acknowledge the duty to file, in this application or patent, notification of any change in status resulting in loss of entitlement to small entity status prior to paying, or at the time of paying, the earliest of the issue fee or any maintenance fee due after the date on which status as a small entity is no longer appropriate. (37 CFR 1.28(b))

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further, that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code, and that such willful false statements may jeopardize the validity of the application, any patent issuing thereon, or any patent to which this verified statement is directed.

Mark W. Perlin

Name of inventor



Date Feb 14, 2000

Signature of Inventor

Name of inventor

Signature of Inventor

Name of inventor

Signature of Inventor

A METHOD AND SYSTEM FOR DNA ANALYSIS

FIELD OF THE INVENTION

5 The present invention pertains to a process for
analyzing a DNA molecule. More specifically, the present
invention is related to performing experiments that produce
quantitative data, and then analyzing these data to characterize a
DNA fragment. The invention also pertains to systems related to
10 this DNA fragment information.

BACKGROUND OF THE INVENTION

With the advent of high-throughput DNA fragment analysis
by electrophoretic separation, many useful genetic assays have
been developed. These assays have application to genotyping,
linkage analysis, genetic association, cancer progression, gene
expression, pharmaceutical development, agricultural improvement,
human identity, and forensic science.

20 However, these assays inherently produce data that have
significant error with respect to the size and concentration of the
characterized DNA fragments. Much calibration is currently done
to help overcome these errors, including the use of in-lane
25 molecular weight size standards. In spite of these improvements,
the variability of these properties (between different
instruments, runs, or lanes) can exceed the desired tolerance of
the assays.

30 Recently, advances have been made in the automated
scoring of genetic data. Many naturally occurring artifacts in
the amplification and separation of nucleic acids can be
eliminated through calibration and mathematical processing of the
data on a computing device (MW Perlin, MB Burks, RC Hoop, and EP

Hoffman, "Toward fully automated genotyping: allele assignment, pedigree construction, phase determination, and recombination detection in Duchenne muscular dystrophy," *Am. J. Hum. Genet.*, vol. 55, no. 4, pp. 777-787, 1994; MW Perlin, G Lancia, and S-K Ng, "Toward fully automated genotyping: genotyping microsatellite markers by deconvolution," *Am. J. Hum. Genet.*, vol. 57, no. 5, pp. 1199-1210, 1995; S-K Ng, "Automating computational molecular genetics: solving the microsatellite genotyping problem," Carnegie Mellon University, Doctoral dissertation CMU-CS-98-105, January 23, 1998), incorporated by reference.

This invention pertains to the novel use of calibrating data and mathematical analyses to computationally eliminate undesirable data artifacts in a nonobvious way. Specifically, the use of allelic ladders and coordinate transformations can help an automated data analysis system better reduce measurement variability to within a desired assay tolerance. This improved reproducibility is useful in that it results in greater accuracy and more complete automation of the genetic assays, often taking less time at a lower cost with fewer people.

Genotyping Technology

Genotyping is the process of determining the alleles at an individual's genetic locus. Such loci can be any inherited DNA sequence in the genome, including protein-encoding genes and polymorphic markers. These markers include short tandem repeat (STR) sequences, single-nucleotide polymorphism (SNP) sequences, restriction fragment length polymorphism (RFLP) sequences, and other DNA sequences that express genetic variation (G Gyapay, J Morissette, A Vignal, C Dib, C Fizames, P Millasseau, S Marc, G Bernardi, M Lathrop, and J Weissenbach, "The 1993-94 Genethon Human Genetic Linkage Map," *Nature Genetics*, vol. 7, no. 2, pp.

246-339, 1994; PW Reed, JL Davies, JB Copeman, ST Bennett, SM Palmer, LE Pritchard, SCL Gough, Y Kawaguchi, HJ Cordell, KM Balfour, SC Jenkins, EE Powell, A Vignal, and JA Todd, "Chromosome-specific microsatellite sets for fluorescence-based, 5 semi-automated genome mapping," *Nature Genet.*, vol. 7, no. 3, pp. 390-395, 1994; L Kruglyak, "The use of a genetic map of biallelic markers in linkage studies," *Nature Genet.*, vol. 17, no. 1, pp. 21-24, 1997; D Wang, J Fan, C Siao, A Berno, P Young, R Sapolsky, G Ghandour, N Perkins, E Winchester, J Spencer, L Kruglyak, L 10 Stein, L Hsie, T Topaloglou, E Hubbell, E Robinson, M Mittmann, M Morris, N Shen, D Kilburn, J Rioux, C Nusbaum, S Rozen, T Hudson, and E Lander, "Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome," *Science*, vol. 280, no. 5366, pp. 1077-82, 1998; P Vos, R Hogers, M Bleeker, 15 M Reijans, T van de Lee, M Hornes, A Frijters, J Pot, J Peleman, M Kuiper, and M Zabeau, "AFLP: a new technique for DNA fingerprinting," *Nucleic Acids Res*, vol. 23, no. 21, pp. 4407-14, 1995; J Sambrook, EF Fritsch, and T Maniatis, *Molecular Cloning*, Second Edition. Plainview, NY: Cold Spring Harbor Press, 1989), 20 incorporated by reference.

The polymorphism assay is typically done by characterizing the length and quantity of DNA from an individual at a marker. For example, STRs are assayed by polymerase chain 25 reaction (PCR) amplification of an individual's STR locus using a labeled PCR primer, followed by size separation of the amplified PCR fragments. Detection of the fragment labels, together with in-lane size standards, generates a signal that permits characterization of the size and quantity of the DNA fragments. 30 From this characterization, the alleles of the STR locus in the individual's genome can be determined (J Weber and P May, "Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction," *Am. J. Hum. Genet.*, vol. 44, pp. 388-396, 1989; JS Ziegler, Y Su, KP Corcoran, L Nie, PE

Mayrand, LB Hoff, LJ McBride, MN Kronick, and SR Diehl, "Application of automated DNA sizing technology for genotyping microsatellite loci," *Genomics*, vol. 14, pp. 1026-1031, 1992), incorporated by reference.

5

The labels can use radioactivity, fluorescence, infrared, or other nonradioactive labeling methods (FM Ausubel, R Brent, RE Kingston, DD Moore, JG Seidman, JA Smith, and K Struhl, ed., *Current Protocols in Molecular Biology*. New York, NY: John Wiley and Sons, 1995; NJ Dracopoli, JL Haines, BR Korf, CC Morton, CE Seidman, JG Seidman, DT Moir, and D Smith, ed., *Current Protocols in Human Genetics*. New York: John Wiley and Sons, 1995; LJ Kricka, ed., *Nonisotopic Probing, Blotting, and Sequencing*, Second Edition. San Diego, CA: Academic Press, 1995), incorporated by reference.

Size separation of fragment molecules is typically done using gel or capillary electrophoresis (CE); newer methods include mass spectrometry and microchannel arrays (RA Mathies and XC Huang, "Capillary array electrophoresis: an approach to high-speed, high-throughput DNA sequencing," *Nature*, vol. 359, pp. 167-169, 1992; KJ Wu, A Stedding, and CH Becker, "Matrix-assisted laser desorption time-of-flight mass spectrometry of oligonucleotides using 3-hydroxypicolinic acid as an ultraviolet-sensitive matrix," *Rapid Commun. Mass Spectrom.*, vol. 7, pp. 142-146, 1993), incorporated by reference.

The label detection method is contingent on both the labels used and the size separation mechanism. For example, with automated DNA sequencers such as the PE Biosystems ABI/377 gel, ABI/310 single capillary or ABI/3700 capillary array instruments, the detection is done by laser scanning of the fluorescently labeled fragments, imaging on a CCD camera, and electronic acquisition of the signals from the CCD camera. Flatbed laser

scanners, such as the Molecular Dynamics Fluorimager or the Hitachi FMBIO/II acquire fluorescent signals similarly. Li-Cor's infrared automated sequencer uses a detection technology modified for the infrared range. Radioactivity can be detected using film 5 or phosphor screens. In mass spectrometry, the atomic mass can be used as a sensitive label. See (A. J. Kostichka, *Bio/Technology*, vol. 10, pp. 78, 1992), incorporated by reference.

Size characterization is done by comparing the sample 10 fragment's signal in the context of the size standards. By separate calibration of the size standards used, the relative molecular size can be inferred. This size is usually only an approximation to the true size in base pair units, since the size standards and the sample fragments generally have different chemistries and electrophoretic migration patterns (S-K Ng, "Automating computational molecular genetics: solving the microsatellite genotyping problem," Carnegie Mellon University, Doctoral dissertation CMU-CS-98-105, January 23, 1998), incorporated by reference.

Quantitation of the DNA signal is usually done by examining peak heights or peak areas. One inexact peak area method simply records the area under the curve; this approach does not account for band overlap between different peaks. It is often 25 useful to determine the quality (e.g., error, accuracy, concordance with expectations) of the size or quantity characterizations. See (DR Richards and MW Perlin, "Quantitative analysis of gel electrophoresis data for automated genotyping applications," Amer. J. Hum. Genet., vol. 57, no. 4 Supplement, 30 pp. A26, 1995), incorporated by reference.

The actual genotyping result depends on the type of genotype, the technology used, and the scoring method. For example, with STR data, following size separation and

characterization, the sizes (exact, rounded, or binned) of the two tallest peaks might be used as the alleles. Alternatively, PCR artifacts (e.g., stutter, relative amplification) can be accounted for in the analysis, and the alleles determined after mathematical 5 corrections have been applied. See (MW Perlin, "Method and system for genotyping," U.S. Patent #5,541,067, Jul. 30, 1996; MW Perlin, "Method and system for genotyping," U.S. Patent #5,580,728, Dec. 3, 1996), incorporated by reference.

10 Genotyping Applications

Genotyping data can be used to determine how mapped markers are shared between related individuals. By correlating this sharing information with phenotypic traits, it is possible to localize a gene associated with that inherited trait. This approach is widely used in genetic linkage and association studies (J Ott, *Analysis of Human Genetic Linkage*, Revised Edition. Baltimore, Maryland: The Johns Hopkins University Press, 1991; N Risch, "Genetic Linkage and Complex Diseases, With Special Reference to Psychiatric Disorders," *Genet. Epidemiol.*, vol. 7, pp. 3-16, 1990; N Risch and K Merikangas, "The future of genetic studies of complex human diseases," *Science*, vol. 273, pp. 1516-1517, 1996), incorporated by reference.

25 Genotyping data can also be used to identify individuals. For example, in forensic science, DNA evidence can connect a suspect to the scene of a crime. DNA databases can provide a repository of such relational information (CP Kimpton, P Gill, A Walton, A Urquhart, ES Milligan, and M Adams, "Automated 30 DNA profiling employing multiplex amplification of short tandem repeat loci," *PCR Meth. Appl.*, vol. 3, pp. 13-22, 1993; JE McEwen, "Forensic DNA data banking by state crime laboratories," *Am. J. Hum. Genet.*, vol. 56, pp. 1487-1492, 1995; K Inman and N Rudin, An

Introduction to Forensic DNA Analysis. Boca Raton, FL: CRC Press, 1997; CJ Fregeau and RM Fournier, "DNA typing with fluorescently tagged short tandem repeats: a sensitive and accurate approach to human identification," *Biotechniques*, vol. 15, no. 1, pp. 100-119, 5 1993), incorporated by reference.

Linked genetic markers can help predict the risk of disease. In monitoring cancer, STRs are used to assess microsatellite instability (MI) and loss of heterozygosity (LOH) - 10 chromosomal alterations that reflect tumor progression. (ID Young, *Introduction to Risk Calculation in Genetic Counselling*. Oxford: Oxford University Press, 1991; L Cawkwell, L Ding, FA Lewis, I Martin, MF Dixon, and P Quirke, "Microsatellite instability in colorectal cancer: improved assessment using 15 fluorescent polymerase chain reaction," *Gastroenterology*, vol. 109, pp. 465-471, 1995; F Canzian, A Salovaara, P Kristo, RB Chadwick, LA Aaltonen, and A de la Chapelle, "Semiautomated assessment of loss of heterozygosity and replication error in tumors," *Cancer Research*, vol. 56, pp. 3331-3337, 1996; S Thibodeau, G Bren, and D Schaid, "Microsatellite instability in cancer of the proximal colon," *Science*, vol. 260, no. 5109, pp. 816-819, 1993), incorporated by reference.

For crop and animal improvement, genetic mapping is a 25 very powerful tool. Genotyping can help identify useful traits of nutritional or economic importance. (HJ Vilkki, DJ de Koning, K Elo, R Velmala, and A Maki-Tanila, "Multiple marker mapping of quantitative trait loci of Finnish dairy cattle by regression," *J. Dairy Sci.*, vol. 80, no. 1, pp. 198-204, 1997; SM Kappes, JW Keele, RT Stone, RA McGraw, TS Sonstegard, TP Smith, NL Lopez-Corrales, and CW Beattie, "A second-generation linkage map of the bovine genome," *Genome Res.*, vol. 7, no. 3, pp. 235-249, 1997; M Georges, D Nielson, M Mackinnon, A Mishra, R Okimoto, AT Pasquino, LS Sargeant, A Sorensen, MR Steele, and X Zhao, "Mapping

quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing," *Genetics*, vol. 139, no. 2, pp. 907-920, 1995; GA Rohrer, LJ Alexander, Z Hu, TP Smith, JW Keele, and CW Beattie, "A comprehensive map of the porcine 5 genome," *Genome Res.*, vol. 6, no. 5, pp. 371-391, 1996; J Hillel, "Map-based quantitative trait locus identification," *Poult. Sci.*, vol. 76, no. 8, pp. 1115-1120, 1997; HH Cheng, "Mapping the chicken genome," *Poult. Sci.*, vol. 76, no. 8, pp. 1101-1107, 1997), incorporated by reference.

10

Other Sizing Assays

Fragment analysis finds application in other genetic methods. Often fragment sizes are used to multiplex many experiments into one shared readout pathway, where size (or size range) serves an index into post-readout demultiplexing. For example, multiple genotypes are typically pooled into a single lane for more efficient readout. Quantifying information can help determine the relative amounts of nucleic acid products present in tissues. (GR Taylor, JS Noble, and RF Mueller, "Automated analysis of multiplex microsatellites," *J. Med. Genet.*, vol. 31, pp. 937-943, 1994; LS Schwartz, J Tarleton, B Popovich, WK Seltzer, and EP Hoffman, "Fluorescent multiplex linkage analysis and carrier detection for Duchenne/Becker muscular dystrophy," *Am. J. Hum. Genet.*, vol. 51, pp. 721-729, 1992; CP Kimpton, P Gill, A Walton, A Urquhart, ES Millican, and M Adams, "Automated DNA profiling employing multiplex amplification of short tandem repeat loci," *PCR Meth. Appl.*, vol. 3, pp. 13-22, 1993), incorporated by reference.

25
30

Differential display is a gene expression assay. It performs a reverse transcriptase PCR (RT-PCR) to capture the state of expressed mRNA molecules into a more robust DNA form. These

DNAs are then size separated, and the size bins provide an index into particular molecules. Variation at a size bin between two tissue assays is interpreted as a concomitant variation in the underlying mRNA gene expression profile. A peak quantification at 5 a bin estimates the underlying mRNA concentration. Comparison of the quantitation of two different samples at the same bin provides a measure of relative up- or down-regulation of gene expression. (SW Jones, D Cai, OS Weislow, and B Esmaeli-Azad, "Generation of multiple mRNA fingerprints using fluorescence-based differential 10 display and an automated DNA sequencer," *BioTechniques*, vol. 22, no. 3, pp. 536-543, 1997; P Liang and A Pardee, "Differential display of eukaryotic messenger RNA by means of the polymerase chain reactions," *Science*, vol. 257, pp. 967-971, 1992; KR 15 Luehrsen, LL Marr, E van der Knaap, and S Cumberledge, "Analysis of differential display RT-PCR products using fluorescent primers and Genescan software," *BioTechniques*, vol. 22, no. 1, pp. 168-174, 1997), incorporated by reference.

Single stranded conformer polymorphism (SSCP) is a 20 method for detecting different mutations in a gene. Single base pair changes can markedly affect fragment mobility of the conformer, and these mobility changes can be detected in a size separation assay. SSCP is of particular use in identifying and diagnosing genetic mutations (M Orita, H Iwahana, H Kanazawa, K 25 Hayashi, and T Sekiya, "Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms," *Proc Natl Acad Sci USA*, vol. 86, pp. 2766-2770, 1989), incorporated by reference.

30 The AFLP technique provides a very powerful DNA fingerprinting technique for DNAs of any origin or complexity. AFLP is based on the selective PCR amplification of restriction fragments from a total digest of genomic DNA. The technique involves three steps: (i) restriction of the DNA and ligation of

oligonucleotide adapters, (ii) selective amplification of sets of restriction fragments, and (iii) gel analysis of the amplified fragments. PCR amplification of restriction fragments is achieved by using the adapter and restriction site sequence as target sites 5 for primer annealing. The selective amplification is achieved by the use of primers that extend into the restriction fragments, amplifying only those fragments in which the primer extensions match the nucleotides flanking the restriction sites. Using this 10 method, sets of restriction fragments may be visualized by PCR without knowledge of nucleotide sequence. The method allows the specific co-amplification of high numbers of restriction fragments. The number of fragments that can be analyzed simultaneously, however, is dependent on the resolution of the detection system. Typically 50-100 restriction fragments are amplified and detected on denaturing polyacrylamide gels. (P Vos, R Hogers, M Bleeker, M Reijans, T van de Lee, M Hornes, A Frijters, J Pot, J Peleman, M Kuiper, and M Zabeau, "AFLP: a new technique for DNA fingerprinting," *Nucleic Acids Res*, vol. 23, no. 21, pp. 4407-14, 1995), incorporated by reference.

Data Scoring

The final step in any fragment assay is scoring the data. This is typically done by having people visually review 25 every experiment. Some systems (e.g., PE Informatics' Genotyper program) perform an initial computer review of the data, to make the manual visual review of every genotype easier. More advanced systems (e.g., Cybergeneitics' TrueAllele technology) fully automate the data review, and provide data quality scores that can 30 be used to identify data artifacts (for eliminating such data from consideration) and rank the data scores (to focus on just the 2%-25% of suspect data calls). See (B Pálsson, F Pálsson, M Perlin, H Gubjartsson, K Stefánsson, and J Gulcher, "Using quality

measures to facilitate allele calling in high-throughput genotyping," *Genome Research*, vol. 9, no. 10, pp. 1002-1012, 1999; MW Perlin, "Method and system for genotyping," U.S. Patent #5,876,933, Mar. 2, 1999), incorporated by reference.

5

However, even with such advanced scoring technology, artifacts can obscure the results. More importantly, insufficient data calibration can preclude the achievement of very low (e.g., <1%) data error rates, regardless of the scoring methods. For 10 example, in high-throughput STR genotyping, differential migration of a sample's PCR fragments relative to the size standards can produce subtle shifts in detected size. This problem is worse when different instruments are used, or when size separation protocols are not entirely uniform. The result is that fragments can be incorrectly assigned to allele bins in a way that cannot be corrected without recourse to additional information (e.g., pedigree data) completely outside the STR sizing assay.

0
10
20
30
40
50
60
70
80
90

Whole System

20 This invention centers on a new way to greatly reduce sizing and quantitation errors in fragment analysis. By designing data generation experiments that include the proper calibration data (e.g., internal lane standards, allelic ladders, uniform run 25 conditions), most of these fragment analysis errors can be eliminated entirely. Moreover, computer software can be devised that fully exploits these data calibrations to automatically identify artifacts and rank the data by quality. The result is a largely error-free system that requires minimal (if any) human 30 intervention.

SUMMARY OF THE INVENTION

The present invention pertains to a method for analyzing a nucleic acid sample. The method comprises the steps of forming 5 labeled DNA sample fragments from a nucleic acid sample. Then there is the step of size separating and detecting said sample fragments to form a sample signal. Then there is the step of forming labeled DNA ladder fragments corresponding to molecular lengths. Then there is the step of size separating and detecting 10 said ladder fragments to form a ladder signal. Then there is the step of transforming the sample signal into length coordinates using the ladder signal. Then there is the step of analyzing the nucleic acid sample signal in length coordinates.

The present invention also pertains to a system for analyzing a nucleic acid sample. The system comprises means for forming labeled DNA sample fragments from a nucleic acid sample. The system further comprises means for size separating and detecting said sample fragments to form a sample signal, said separating and detecting means in communication with the sample fragments. The system further comprises means for forming labeled DNA ladder fragments corresponding to molecular lengths. The system further comprises means for size separating and detecting 25 said ladder fragments to form a ladder signal, said separating and detecting means in communication with the ladder fragments. The system further comprises means for transforming the sample signal into length coordinates using the ladder signal, said transforming means in communication with the signals. The system further comprises means for analyzing the nucleic acid sample signal in 30 length coordinates, said analyzing means in communication with the transforming means.

The present invention also pertains to a method for generating revenue from computer scoring of genetic data. The

method comprises the steps of supplying a software program that automatically scores genetic data. Then there is the step of forming genetic data that can be scored by the software program. Then there is the step of scoring the genetic data using the 5 software program to form a quantity of genetic data. Then there is the step of generating a revenue from computer scoring of genetic data that is related to the quantity.

10 The present invention also pertains to a method for producing a nucleic acid analysis. The method comprises the steps of analyzing a first nucleic acid sample on a first size separation instrument to form a first signal. Then there is the step of analyzing a second nucleic acid sample on a second size separation instrument to form a second signal. Then there is the 15 step of comparing the first signal with the second signal in a computing device with memory to form a comparison. Then there is the step of producing a nucleic acid analysis of the two samples from the comparison that is independent of the size separation instruments used.

20 The present invention also pertains to a method for resolving DNA mixtures. The method comprises the steps of obtaining DNA profile data that include a mixed sample. Then there is the step of representing the data in a linear equation. 25 Then there is the step of deriving a solution from the linear equation. Then there is the step of resolving the DNA mixture from the solution.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 shows the steps of creating sized profiles.

5 Figure 2 shows unimodal plots, each corresponding to a fluorescent dye; within each plot, intensity is plotted against the sampled spectrum.

10 Figure 3 shows the unimodality constraint determining the function space geometry of the spectrum sampling vectors.

Figure 4 shows the results of signal processing, size tracking, and size transformation.

15 Figure 5 shows the steps of quantitating and analyzing genetic data.

Figure 6 shows the results of ladder processing, peak quantitation and allele calling.

20 Figure 7 shows a graphical user interface for navigating prioritized genotyping results.

Figure 8 shows a textual interface for displaying useful genotype results.

25 Figure 9 shows a visualization that is customized to a data artifact.

30 Figure 10 shows a system for analyzing a nucleic acid sample.

Figure 11 shows the result of a differential display gene expression analysis.

Figure 12 shows the flow graph of automated software assembly.

Figure 13 shows a spreadsheet for calculating the labor cost of scoring genetic data.

5

Figure 14 shows a heuristic function $dev(\mathbf{g}(\mathbf{w}))$ which has an unambiguous local minimum.

DESCRIPTION OF THE PREFERRED EMBODIMENT

Data Generation

5 In the most preferred embodiment, genotyping data is generated using STR markers. These tandem repeats include mono-, di-, tri-, tetra-, penta-, hexa-, hepta-, octa-, nona-, deca- (and so on) nucleotide repeat elements. STRs are highly abundant and informative marker distributed throughout the genomes of many
10 species (including human). Typically, STRs are labeled, PCR amplified, and then detected (for size and quantity) on an electrophoretic gel.

0
15
20
25
30

The laboratory processing starts with the acquisition of a sample, and the extraction of its DNA. The extraction and purification are typically followed by PCR amplification. Labelling is generally done using a 5' labeled PCR primer, or with incorporation labeling in the PCR. Prior to loading, multiple marker PCR products in $k-1$ different fluorescent colors are pooled, and size standards (preferably in a k^{th} different color) is added. Size separation and detection is preferably done using automated fluorescent DNA sequencers, with either slab gel or capillary technology. The detected signals represent the progression of DNA bands as a function of time. These signals are
25 transmitted from the sequencer to a computing device with memory, where they are stored in a file. (NJ Dracopoli, JL Haines, BR Korf, CC Morton, CE Seidman, JG Seidman, DT Moir, and D Smith, ed., Current Protocols in Human Genetics. New York: John Wiley and Sons, 1995), incorporated by reference.

30

To create STR allelic ladders, the most preferred embodiment entails PCR amplification of pooled samples. This can be done by preparing DNA from N (preferably, N is between 2 and

200, depending on the application) individuals in equimolar 3 ng/ μ l concentrations. These DNAs are then pooled. After dilution, each PCR template contains 48 ng of DNA in an 18 μ l volume, and is included in a standard 50 μ l PCR containing 5 2.5 units of AmpliTaq Gold, 1.25 μ l of each primer, 200uM dNTP's, and 2.5mM MgCl₂. This mixture is then PCR amplified with its STR primers (one labeled) on a thermocycler (e.g., with an MJ Research PTC-100, use 30 cycles of 94°C for 1.25', 55°C for 1', and 72°C for 1'). Size separation of the PCR products on an ABI sequencer 10 includes internal lane size standards (GS500 ROX-labeled 50 bp sizing ladder, Perkin-Elmer, Foster City, CA; 20 bp MapMarkers sizing ladder, BioVentures, Murfreesboro, TN). Files are similarly recorded from this experiment.

15 In an alternative preferred embodiment, multiplexed SNP data is generated using size standards with standard protocols. Typically, each size bin in the electrophoretic signal corresponds to one marker or polymorphism. Presence in the size bin of a signal of sufficient strength and correct color indicates the presence of an allele; absence of the signal indicates allele 20 absence. Signal size and color establish the allele, while signal strength determines the amount of DNA present (if any). (NJ Dracopoli, JL Haines, BR Korf, CC Morton, CE Seidman, JG Seidman, DT Moir, and D Smith, ed., *Current Protocols in Human Genetics*. 25 New York: John Wiley and Sons, 1995), incorporated by reference.

In another alternative preferred embodiment, differential display data is generated, preferably as follows:

Making cDNA. The mRNA differential display reverse 30 transcription-polymerase chain reaction (DDRT-PCR) is performed using reagents supplied in a RNAimage™ kit (GeneHunter Corp, Nashville, TN). RNA duplicates from two tissue samples are reverse-transcribed using oligo(dT) primers and MMLV reverse transcriptase (RTase). For a 20 μ l reaction, adding (in order)

9.4 ul of H₂O, 4 ul of 5x reaction buffer, 1.6 ul of 250 uM each dNTPs, 2 ul of 2 uM one oligo(dT) primer (H-T11M: M is G, C, or A), and 2 ul of 0.1 ug/ul DNA-free total RNA. The RT reactions are run on a thermocycler (e.g., MJ Research, 65°C for 5 min, 37°C 5 for 60 min, and 75°C for 5 min). MMLV RTase (1 ul, 100 units) are added in the reaction after incubation for 10 min at 37°C. A control is included without adding RTase.

Amplification and labeling. For a 20 ul PCR reaction, add 9.2 ul of H₂O, 4 ul of 10x PCR buffer, 1.6 ul of 25 uM each dNTPs, 2 ul 10 of 2 uM H-T11M oligo(dT) primer, 2 ul of 2 uM of arbitrary primer (AP), 2 ul of corresponding H-T11M reverse transcribed cDNA, and 0.2 ul (1 unit) of AmpliTaq DNA polymerase (Perkin Elmer, Norwalk, CT). Each of the three different H-T11M PCR primers are labeled with its own spectrally distinct fluorescent dye, such as FAM, 15 HEX, and NED. The cDNAs are randomly amplified by low-stringency PCR (40 cycles with temperature at 94°C for 15 sec, 40°C for 2 min, and 72°C for 2 min) in an MJR PCT/100 thermocycler. A final extension is performed at 72°C for 10 min. Samples (without added 20 RTase or cDNA) are simultaneously tested as controls. Multiple primer sets can be used. For example, 24 sets of primers (8 AP x 3 H-T11M) are used in each kit; using 10 kits for screening differentially expressed cDNA tags produces 240 reactions per tissue.

Size separation. Size standards in another dye are then added 25 to the amplified labeled products, and then size separated on a manual or automated sequencing gel (or capillary) instrument. Differential display data generation protocols have been well described (NJ Dracopoli, JL Haines, BR Korf, CC Morton, CE Seidman, JG Seidman, DT Moir, and D Smith, ed., Current Protocols 30 in Human Genetics. New York: John Wiley and Sons, 1995), incorporated by reference.

There are other alternative preferred embodiments for generating DNA fragment data whose assay includes a size separation, such as Amplification Refractory Mutation System 5 (ARMS), Single-Strand Conformation Polymorphism (SSCP), Restriction Fragment Length Polymorphism (RFLP), and Amplified Fragment Length Polymorphism (AFLP). These have been enumerated, with associated protocols (NJ Dracopoli, JL Haines, BR Korf, CC Morton, CE Seidman, JG Seidman, DT Moir, and D Smith, ed., Current 10 Protocols in Human Genetics. New York: John Wiley and Sons, 1995; ABI/377 and ABI/310 GeneScan Software and Operation Manuals, PE Biosystems, Foster City, CA), incorporated by reference.

Extracting Profiles

Once DNA fragment sizing data have been generated, the data are then analyzed to characterize the and quantity of the component fragments.

20 Referring to Figure 1, Step 1 is for acquiring the data.

The process begins by reading in the generated data from their native file formats, as defined by the DNA sequencer manufacturer. Let n be the number of lanes or capillaries, and m 25 be the number of frequency acquisition channels. Capillary machines typically produce files where each file represents either all m channels of one capillary, or one channel of one capillary. Gel-based instruments typically produce files where one file represents either all m channels of the entire image, or one 30 channel of one image. Intermediate cases, with the number of channels per file between 1 and m , can occur as well.

Once the signals have been read in, capillary input data signals are preferably stored in an $n \times m$ structure of one dimensional arrays in the memory of a computing device. This structure contains the signal profiles, with each array element 5 corresponding to one channel of one capillary. Gel data are preferably stored as m two dimensional data arrays, one for each acquisition frequency.

The computer software preferably integrates with current 10 sequencer and CE technology. It preferably has two manufacturer-independent input modules: one for sequencer gel data (e.g., PE Biosystems ABI/377, Molecular Dynamics Fluorimager), and one for CE data (e.g., PE Biosystems ABI/310, SpectruMedix SCE/9600). These modules are extensible and flexible, and preferably handle 15 any known sequencer or CE data in current (or future) file formats.

Referring to Figure 1, Step 2 is for processing the signal.

In this step, basic signal processing is done, such as 20 baseline removal or filtering (e.g., smoothing) the data.

In the preferred embodiment for one dimensional signals, baseline removal is done using a sliding window technique. Within 25 each window (of, say, 10 to 250 pixels, depending on the average number of pixel samples per base pair units), a minimum value is identified. Using overlapping windows, a cubic spline is fit through the minimum points, creating a new baseline function that describes the local minima in each window neighborhood. To remove 30 the baseline, this new baseline function is subtracted away from the original function.

In the preferred embodiment for two dimensional images, the baseline is removed as follows. A local neighborhood overlapping tiling is imposed on the image, and minimum values identified. Create a baseline surface from these local minima, 5 and subtract this baseline surface from the image to remove the baseline from the image.

In the preferred embodiment, filtering and other smoothing is done using convolution. A convolution kernel (such 10 as a gaussian or a binomial function) is applied across the one dimensional capillary signal, or the two dimensional scanned gel image. The radius of smoothing depends on number of pixels per base pair unit - denser sampling requires less smoothing. 15 However, with overly dense sampling, the data size can be reduced by filtering out redundant points.

Referring to Figure 1, Step 3 is for separating the colors.

A key element of fluorescent genetic analysis is separating the fluorescent dye signals. In the current art, this 20 is done by:

- (1) Performing a dye standard calibration experiment using known dyes, often in separate lanes. A (dye color vs. frequency detection) classification matrix **C** is known 25 directly from which dye is used. Each column of **C** contains a "1" in the row corresponding to the known color, with all other entries set to "0".
- (2) Measuring the signals at separate fluorescent detection frequencies. These frequencies correspond to the "filter set" of the sequencer. For each pure dye peak, the signals 30 that are measured across all the detection frequencies reveal the "bleedthrough" pattern of one dye color into its neighboring frequencies. Each pattern is normalized, and stored as a column in a data matrix **D**. The j^{th} column of **D**

is the "spectral bleedthrough" system response to the impulse input function represented in the j th column of \mathbf{c} .

(3) The relationship between **C** and **D** is described by the linear response dye calibration matrix **M** as:

$$\mathbf{D} = \mathbf{M} \times \mathbf{C}$$

To determine the unknown dye calibration matrix **M** (or its inverse matrix \mathbf{M}^{-1}) from the data, apply matrix division, e.g., using singular value decomposition (SVD), to the matrices **C** and **D**. For example, this SVD operation is built into the MATLAB programming language (The MathWorks, Inc., Natick, MA), and has been described (WH Press, SA Teukolsky, WT Vetterling, and BP Flannery, Numerical Recipes in C: The Art of Scientific Computing, Second Edition. Cambridge: Cambridge University Press, 1992), incorporated by reference.

(4) Thereafter, the spectral overlap is deconvolved on new unknown data \mathbf{D}' to recover the original dye colors \mathbf{C}' . This is done by computing:

$$\mathbf{C}' = \mathbf{M}^{-1} \times \mathbf{D}'$$

While the current art enables the color separation step, it is not ideal. Slight variations in the gel (e.g., thickness, composition, temperature, chemistry) and the detection unit (e.g., laser, CCD, optics) can contribute to larger variations in fluorescent response. An operator may encounter the following problems:

- Calibrating the correction matrix \mathbf{M}^{-1} on one gel run does not necessarily model the "spectral bleedthrough" pattern accurately on future runs.
- With 96 (or other multi-) capillary electrophoresis, each of the capillaries forms its own gel system, whose different properties may necessitate a separate calibration matrix.

- Accurate dye calibration is technically demanding, labor intensive, time consuming, and expensive. Moreover, it can introduce considerable error into the system, particularly when the manual procedure is not carried out correctly. In such cases, the correction is imperfect and artifacts can enter the system.

Such color "bleedthrough" (also termed "crosstalk" or "pullup") artifacts can severely compromise the utility of the acquired data. In some cases, the gels must be rerun. Often scientific personnel waste considerable time examining highly uninformative data.

In one preferred embodiment, the color matrix is calibrated directly from the data, without recourse to separate calibration runs. This bleedthrough artifact removal is done using computer algorithms for the calibration, rather than manually conducting additional calibration experiments. This can be done using methods developed for DNA sequence analysis based on general data clustering. While such clustering methods require relatively large amounts of data, they can be effective (W Huang, Z Yin, D Fuhrmann, D States, and L Thomas Jr, "A method to determine the filter matrix in four-dye fluorescence-based DNA sequencing," *Electrophoresis*, vol. 18, no. 1, pp. 23-5, 1997; Z Yin, J Severin, M Giddings, W Huang, M Westphall, and L Smith, "Automatic matrix determination in four dye fluorescence-based DNA sequencing," *Electrophoresis*, vol. 17, no. 6, pp. 1143-50, 1996), incorporated by reference.

It would be desirable to use the least amount of most certain data when determining a color matrix for separating dye colors. Note that in the matrix relation " $\mathbf{D}=\mathbf{M}\times\mathbf{C}$ ", the data columns \mathbf{D} are known from experiment. When a calibration run is used, the 0/1 classification columns \mathbf{C} are known; from these known

values, the unknown \mathbf{M} can be computed. However, when there is no calibration run, the classification matrix \mathbf{C} must be dynamically determined from the data in order to compute \mathbf{M} . This can be done manually by a user identifying peaks, or automatically by the 5 general clustering embodiment. However, there is a more refined and novel approach to automatically classifying certain data peaks to their correct color.

The most preferred embodiment finds a $\{\mathbf{C}, \mathbf{D}\}$ matrix pair 10 using minimal data. Thus, \mathbf{M} (and \mathbf{M}^{-1}) can be determined, and the rest of the new data color separated. This embodiment exploits a key physical fact about spectral emission curves: they are 15 *unimodal*. Referring to Figure 2, such curves monotonically rise up to a maximum value, and then monotonically decrease from that maximum. Therefore, when a peak is generated from a single dye 20 color, spectrally sampled across a range of frequencies, and the intensities plotted as a function of frequency, the plotted curve demonstrates unimodal behavior - the curve rises up to a maximum intensity, and then decreases again. This physical "unimodality" 25 constraint on single color peaks can serve as a useful choice function for automatically (and intelligently) choosing peak data for the classification matrix \mathbf{C} .

The function space equivalent of a unimodal function has 25 a very useful property. Suppose that m different frequency channels are sampled. That is, there are m frequencies $\mathbf{x} = \{x_1, x_2, \dots, x_m\}$ sampled in the spectral domain. An equivalent representation of the m -point function $f: \underline{m} \rightarrow \mathbb{R}$ is as an m -vector $\mathbf{v} = \langle x_1, x_2, \dots, x_m \rangle$ in the vector space \mathbb{R}^m . Now, select an 30 appropriate norm on this vector space (say, L_1 or L_2), and normalize the vector \mathbf{v} relative to its length, forming

$$\mathbf{w} = \frac{\mathbf{v}}{\|\mathbf{v}\|}.$$

In an L_1 normalization, \mathbf{w} lives on a flat simplex (in the all-positive simplex facet). In L_2 , \mathbf{w} lives on a corresponding curved surface.

5

The unimodality constraint on \mathbf{x} imposes additional geometrical constraints on \mathbf{w} . Because \mathbf{x} is unimodal, note that

$$x_1 \leq x_2 \leq \dots \leq x_k,$$

and that

10

$$x_k \geq x_{k+1} \geq \dots \geq x_m,$$

where k is the index of the maximum value of \mathbf{x} .

These inequality constraints determine the exact subfacet in which \mathbf{w} must reside on the simplex facet. Consider the case of three spectral sampling points, where the three-dimensional vector $\langle x, y, z \rangle \in \mathbb{R}^3$, and suppose that the unimodal constraint is $x \geq y \geq z$, corresponding to the first dye color. Referring to Figure 3, the locations x , y , and z designate the unit locations on the axes in \mathbb{R}^3 (i.e., at $\langle 1, 0, 0 \rangle$, $\langle 0, 1, 0 \rangle$, and $\langle 0, 0, 1 \rangle$, respectively). Here, the $x \geq y$ constraint produces one region (horizontal shading), the $y \geq z$ constraint another (vertical shading), and their intersection corresponds to the full constraint (crosshatch shading). Good calibration points for use as columns in a $\{\mathbf{C}, \mathbf{D}\}$ matrix pair will cluster (white circle) around the dye's actual sampling point ratio vector (black cross). Conversely, poor candidate calibration points will tend to lie outside the cluster, and can be rejected. This geometry in the function space permits selection of only the most consistent calibration data.

30

The procedure starts by gathering likely unimodal data (e.g., those with highest intensities) for a given observed color k . After normalization, these candidate calibration data $\{w_i\}$ will cluster within the same subfacet of a flat $(m-1)$ -dimensional 5 facet; this facet is the all-positive face of the m -dimensional simplex. This geometric constraint follows directly from the physical unimodality constraint of pure spectral curves. An entire class of simple and effective clustering algorithms are based around exploiting this geometric constraint.

10

In one preferred embodiment, choose as cluster center point w_0 , the mean vector location of $\{w_i\}$; vector w_0 also lies on the simplex subfacet (Figure 3). Then, a small inner product $\langle w_i, w_0 \rangle$ value tends to indicate close proximity of w_i and w_0 . 15 Taking a small set (e.g., $1 \leq s \leq 20$) of the closest vectors w_i near w_0 selects good calibration data points, all pre-classified to color k . To determine M^{-1} , for each color k , take at most s such clustering points $\{w_i\}$, and use the corresponding $\{v_i\}$ as columns in D , where v_i is normalized with respect to the maximum element 20 in w_i . Form a vector u that is all zeros, except for a 1 in the k^{th} entry; place s copies of u as columns in matrix C . This produces the required calibration matrices C and D , from which M 25 and its inverse are immediately computed using SVD or another matrix inversion algorithm.

25

In the most preferred color separation algorithm, the criterion for peak selection is based on minimizing the spectral width of the peaks. A good measure of peak width is the variance across the spectral sampling frequency points. The variance 30 calculation can take into account the sampling frequencies actually used, if desired. In this procedure, the variances of candidate peaks for a given color frequency are computed. Those peaks having the smallest spectral variance indicate the best

calibration points. In an alternative implementation, a best fit of the observed frequency curve with a known frequency curve (e.g., via a correlation or inner product maximization) can indicate the best points to use.

5

The method applies not only to raw data, but also to data that has been previously color separated by other (possibly inaccurate) color correction matrices. This is because the unimodality constraint generally applies to such data.

10

A useful feature of the unimodality approach is that the model can automatically select good calibration peaks, even with very sparse data. This is because the function space geometry effectively constrains the clustering geometry. Such sparse-data clustering algorithms are particularly useful with capillary data, where one capillary may only have 1 or 2 useful calibration peaks corresponding to a given color. Despite this data limitation, the method easily finds these peaks, and effectively separates the colors. Such automation enables customization of the dye separation matrix to each capillary on each run, which virtually eliminates the bleedthrough artifact.

It is useful to have a quality score that measures how well the computed matrix correction actually corrects the data.

25 This can be done by comparing the expected vs. observed results; this comparison can be computed either the separated or unseparated domain. Using the computer-selected calibration data vectors **D**, if the computed correction matrix \mathbf{M}^{-1} is correct, then the 0/1 calibration matrix of column vectors **C** will be recomputed exactly from the data matrix **D**:

$$\mathbf{C}' = \mathbf{M}^{-1} \times \mathbf{D}$$

where, theoretically, $\mathbf{C}' = \mathbf{C}$. Measuring the deviation between **C** and **C'** measures how well the correction worked. One

straightforward deviation measure sums the normed L_2 deviations between each column of \mathbf{C}' and its corresponding column in \mathbf{C} . When the deviation is too great (e.g., due to failed PCR amplification), a matrix based on a more confident (manual or 5 automatic) calibration can be used.

Referring to Figure 1, Step 4 is for removing the primers.

When the separated DNA fragments are formed from labeled 10 PCR primers, it is useful to remove the intense primer signal prior to initiating quantitative analyses. There are many standard signal processing methods for removing a singularity from a signal, or a row of such large peaks from an image.

For one dimensional data, first detect the primer signal. This can be done, in one embodiment, by smoothing (i.e., low pass filtering) the signal to focus on broad variations, and then fitting the largest peak (i.e., the primer peak) to an appropriate function, such as a Gaussian. Determining the variance of the peak from the function fit provides the domain interval on which the primer peak should be removed. On this domain, it is most preferable to set the values on this interval to an appropriate background value (e.g., zero, if background subtracted, or an average of neighboring values outside the 25 interval). Alternatively, one can crop the interval from the signal.

For two dimensional data, a projection of the pixel data onto the vertical axis of DNA separation finds the row of peaks. 30 Determine the spread of this peak signal by curve fitting (e.g., with a Gaussian). Remove the primer peaks by either cropping the signal from the image, or setting the values in that domain to zero (or some other appropriate background value).

Referring to Figure 1, Step 5 is for tracking the sizes.

In the preferred embodiment, size standards are run in the same lane as the sample data, and are labeled with a label 5 different from the label used for the sample data. The task is to find these size standards peaks, confirm which peaks represent good size standard data, and then align the observed size standards peaks with the expected sizes of the size standards. This process creates a mapping between the size standard peak data 10 sampled in the pixel domain, and the known sizes (say, in base pair or molecular weight units) of each peak.

For one dimensional data, there are no lateral lanes to help determine which of the peaks observed in the size standard signal represent good data. Therefore, a preferred procedure uses prior information about the size standards (e.g., the size) to ensure a proper matching of data peaks to known sizes. In the most preferred embodiment, use the following steps:

0. Find some good candidate peaks to get started.
1. Identify the best peaks to use by filtering poor candidates. This can be done by performing quality checks (e.g., for the height, width, or peak fit) on the candidate peaks.
2. Match the expected peak locations to the observed data peaks. This is done by applying a "zipper match" algorithm to the 25 best candidate data peaks and the expected sizes. This matching uses local extension to align the peaks with sizes, and includes the following steps:
 - 2a. Match the boundary, that is, a subset of smallest size data peaks or largest size data peaks. The algorithm can be rerun several times, shifting the boundary data peaks or the expected sizes. The best boundary shift can be found by heuristic minimization. One good heuristic assesses the uniformity (e.g., by minimizing the variance) of the ratios 30

across matching local intervals of the expected size standard difference to the observed data peak difference.

2b. Fix a tolerance interval that includes unity (different tolerance intervals can be tried).

5 2c. Starting with the (possibly truncated) boundaries of the expected sizes and observed peaks aligned, extend this boundary.

10 2d. Compute the ratio p of the difference between the next expected peak size and the current expected peak size, to the difference between the next observed peak size and the current observed peak size.

2e. Compute the ratio q of the difference between the current expected peak size and the previous expected peak size, to the difference between the current observed peak size and the previous observed peak size.

2f. If the ratio (p/q) is greater than the tolerance interval's greatest value, then the observed data peak falls short. Reject the observed data peak, advance to the next observed data peak, and continue.

2g. If the ratio (p/q) is less than the tolerance interval's least value, then the expected data size falls short. Reject the expected data size, advance to the next expected data size, and continue.

25 2h. If the ratio (p/q) lies within the tolerance interval, then the expected data size is well matched to the observed data peak. Accept the size and the peak, record their match, and advance to the next size and peak.

3. If desired, fill in any missing data peaks by interpolation with expected sizes from the zipper matching result.

30 4. Compute a quality score for the matching result. Useful scores include:

4a. Fitting the matching of the expected sizes (in base pair, or other expected unit) with the peak sizes (in pixels, or other observed data unit). The relationship is monotonic

and typically slowly varying, so a deviation from a fitted function (e.g., linear, cubic, or Southern mobility relationship) works well.

4b. Another quality score is the number of size mismatches,
5 adjusted by the boundary shift.

5. Report the peak positions, matched with size standards. It is also useful to include the quality score of the match.

10 For two dimensional gel data, proceed by tracking on the color-separated size standard image. To track simultaneously both the sizes and lanes, first focus on a boundary row (e.g., the top row) of size standards. Model the row geometry (e.g., curvature, size, location) and each of the peaks (e.g., height, shape, angle). Then use this pattern as a set of geometrical constraints for analyzing the next row. Use global search to find the next row, and more localized search to identify the peaks; refine the local peak locations using center-of-mass calculations. Continue this process iteratively until the size standard data are completely analyzed, and a final grid is produced. It is possible to use a prior analysis of the loading/run pattern as calibration data in order to speed up subsequent tracking. The output of this lane/size tracking is a mapping between the expected (lane, base pair size) coordinate grid, and the observed (x pixel, y pixel) gel data image coordinate grid.

25

30 The quality of the tracking result can be scored by comparing the expected grid locations with the observed data peak locations - straight data grids indicate a high quality result, whereas large distortions indicate a lower quality result. One useful quality score measures the curvature of the observed peaks by forming the sum of local neighbor distances, and normalizing relative to the neighbor distances in a straight idealized grid. This grid is preferably formed from the known lane loading

pattern, the known size standard positions, and observed grid data (e.g., mean boundary positions).

With two dimensional gel image data $f(x,y,z)$, z denoting the color plane, there also is the step of extracting lane profiles from the gel image to form a set of one dimensional profiles. The lane tracking result is used in this step. For each lane, form the mapping from y -pixel values to the (x,y) pixel locations indicated by the size standard location found for the lane. Create a spline function (e.g., cubic) that interpolates through the mapped $y \rightarrow x$ pixel pairs to all other (x,y) pixel values. Then, extract a one dimensional profile (for each color k), where the domain values down the lane are the desired sequential pixels y , and the range values are computed from the function $f(v(x),y,z)$. This procedure is done for each data image plane z , extracting either color separated data (by applying the correction matrix) or unseparated data.

Referring to Figure 1, Step 6 is for extracting the profiles.

The signal profile of a capillary or lane is $f(x_{pixel},z)$, where z denotes the label color (or channel). This function is available for capillary data available after processing Step 2, and for gel data after extracting profiles in Step 5.

The pixel sampling of the electrophoresis distorts the sizing of the raw data. The size tracking result from Step 5 provides a set of (x_{pixel}, x_{size}) pairs. Use these matched pairs to form the coordinate transformation $\mu: x_{pixel} \rightarrow x_{size}$. Combining the functions f and μ via a double interpolation to form:

$$f \circ \mu^{-1}(x_{size}, z),$$

a new function that describes the signal as a function of size standard units (instead of pixel units), preferably in base pair size units. If color separation has not yet been done, the correction matrix is applied at this time. These transformed 5 (capillary or gel data) profiles are then preferably stored in an $n \times k$ data structure of sized profiles, where n is the number of pixel samples, and k is the number of dye colors.

Note that the $f_{pixel} \circ \mu^{-1}(x_{size}, z)$ transformation which 10 maps

$$x_{size} \rightarrow f_{size}(x_{size})$$

can be usefully understood as the commutative diagram:

$$\begin{array}{ccc} x_{size} & \xrightarrow{\quad} & f_{size}(x_{size}) \\ | & & \uparrow f_{size} \\ | \mu^{-1} & & | \text{ identity} \\ \downarrow & & \downarrow f_{pixel} \\ x_{pixel} & \xrightarrow{\quad} & f_{pixel}(x_{pixel}) \end{array}$$

The transformation is preferably implemented by 20 performing a double interpolation. In MATLAB, this operation can be readily computed using the 'interp1' function where:

`YI = interp1(X, Y, XI, method)`

interpolates to find YI , the interpolated values of the underlying function Y at the points in the vector XI ; the vector X specifies 25 the points at which the data Y is given.

With 'interp1', interpolate the desired size *domain* points into computed *pixels* using the monotonic bijection between expected *sizes* and observed *peak* pixels:

30 `pixels = interp1(sizes, peaks, domain, 'spline')`

Cubic spline interpolation is done by setting the method to 'spline'. Then, interpolate the function f on the desired size

domain points (i.e., the computed *pixels*) using the monotonic bijection $f(i)$ between *indices* { i } and *data* values $f\{i\}$:

```
indices = [1:length(data)];  
profile = interp1(indices, data, pixels, 'spline');
```

5 Computing *profile* completes the commutative diagram, and produces the new function $f_{size}(x_{size})$.

Referring to Figure 4, visualizations are shown for the results of processing Steps 1-6.

10

The computer software preferably extracts and analyzes data automatically, without human intervention (if so desired). The software separates colors using a matrix, which is either precalibrated or created from the data, depending on which module is used. The software tracks lanes and size standards on two-dimensional gel data automatically by mapping the expected two dimensional lane/size grid to the observed size standard data; on one-dimensional CE data, size tracking is done separately for each capillary. The user interface software makes manual retracking, zooming, and single-click access to the chemistry panel (or sample data used on a run) available to the user throughout.

25 A chain of custody is maintained in that the user preferably cannot move to the next step without saving results (manually or automatically) from the current step. Changes are saved, not discarded; moreover, the software records these changes incrementally, so that the audit trail cannot be lost by early program termination.

30 Backtracking capability and flexibility are preferably included in the software. For example, should fully automated lane tracking fail due to low-quality data, the user can choose to: (a) edit the results, and have the program re-track, (b) edit the results without re-tracking, (c) manually track the lanes and

sizes, or (d) reject the low-quality gel. A "revert" operation provides a universal Undo operation for automatically rolling back major processing steps.

5 Data Scoring

Referring to Figure 5, Step 7 is for deriving an allelic ladder.

10 It is useful to have a set of reference peaks that (a) correspond to the actual locations of DNA molecules on the gel, (b) have known lengths (in base pair units), and (c) cover a large part of the sizing window. This reference set can be developed for any fragment sizing genetic assay; without loss of generality, the preferred embodiment is described for STR genotyping.

20 Construct a partial allelic ladder by PCR amplifying a pool of DNA samples. This allelic ladder, and optionally a known sample, are preferably loaded into the electrophoretic system as separate signals (i.e., in different lanes or colors). After the gel is run, referring to Figure 1, Steps 1-6 are performed to initially process the signals. Then, preferably in size coordinates (rather than in pixel coordinates), a peak finding algorithm locates and sizes the peaks in these two lanes with respect their in-lane size standards. The known sample's sized 25 peaks are then compared with the similarly sized peaks found in the ladder lane. Following this comparison, the known DNA lengths are then assigned to the allelic ladder. The DNA length labels of these DNA lengths are then propagated to the unlabeled ladder peaks.

30

A preferred peak labeling procedure is:

1. In the ladder signal, find the domain positions (i.e., the sizes relative to the size standards) of the allelic peaks.

2. Perform a relational labeling from the known signal 5 to the allelic ladder signal, as follows. Find the peaks of the known signal, and assign to them their known sizes (in allelic sizes units, such as base pairs). Then, match (in size standard units) these known allele peaks to the peaks in the ladder signal having corresponding size. Designate these ladder peaks with the 10 allelic size labels of the known peaks.

3. Extend these allelic label assignments from the allelic ladder peaks designated with known labels to the rest of the ladder peaks.

3a. When the expected size pattern of the alleles on the allelic ladder is known (as with previously characterized forensic ladders), a robust method for assigning size labels to data peaks uses a standard open/closed search set artificial intelligence (AI) algorithm. Start from the most confident data (i.e., the knowns) as the closed set, with all remaining peaks in the open set. Each cycle, (i) select an open ladder allele that is nearest 20 in allele size to a closed ladder allele, (ii) predict the calibrated size (from the size standards) of this allele using interpolation relative to the closed peaks, (iii) find the allelic ladder peak whose calibrated size is closest to the predicted size 25 in (ii), and (iv) if the quality score (based on size deviation) is good enough, move this candidate open peak to the closed set. On termination, only the most confident ladder peaks in the observed data have been matched to expected ladder alleles. See (NJ Nilsson, Principles of Artificial Intelligence. Palo Alto, CA: 30 Tioga Publishing Co., 1980), incorporated by reference.

3b. When expected ladder alleles are not available (e.g., with uncharacterized pooled DNA samples), the ladder pattern alone (peak spacing or peak heights) usually contains sufficient information for designating the labeling. Start with

the most confident data - known allele peaks, or tallest ladder peaks. Then, locally extend the allele labels to the size peaks. Since there is more uncertainty without a known ladder pattern, more search is useful. One preferred method is to (i) find the 5 tallest (most confident) peak, (ii) match size with nearest (bp size) allele, (iii) locally align iteratively for smaller and larger peaks, (iv) assign a quality score to the alignment, (v) repeat the preceding three steps, but shifted one or two sizes up or down, and (vi) select the alignment with the highest quality 10 score.

4. Optionally, fill in any missing ladder peaks with interpolation by reference to past ladder data.

5. Fill in the virtual points on the ladder. These are expected ladder alleles that are believed to exist in the population, but that are not represented as peaks in the allelic ladder. This is done by interpolation.

6. Return the results assigning allele labels (e.g., in base pair or other integer units) to data ladder peaks (e.g., in size standard units). It is preferable to report the quality score of the assignment.

In nonforensic STR applications, where previously characterized ladders are generally unavailable, pooled alleles of a given marker can be used as a reference ladder. This novel 25 approach can help eliminate the size binning problem that plagues microsatellite and other STR genetic methods. It is preferable to use the same allelic ladder across multiple runs. The ladder can be comprised of either pooled DNAs that are PCR amplified together, or post-PCR amplified products that are then pooled 30 together. It may be desirable to visually inspect the ladders. In one preferred embodiment, previously uncharacterized ladders are checked the first time that they are encountered, with a human editor identifying the best peaks to use and matching them against their expected sizes. Recording quantitative peak data for a

ladder can enable the use of quantitative computer-based matching of reference ladders and new unknown ladders (e.g., for peak and size alignment) by correlation or other inner product methods.

5 When the interpolation and extrapolation have finished, the allele sizes that were actually present as peaks in the ladder data, as well as desired allele sizes that were not present as peaks, have all been allocated to size positions.

10 SNP ladders can be developed in a similar way to STR ladders. With multiplexed SNPs, it is useful to run allelic sizing ladders comprised of actual SNP data for the markers of interest in a signal. Then, comparison of the unknown SNP sample data with the SNP allelic ladder can remove uncertainty regarding the correct size (hence, allele) assignment. The same reasoning applies to any size-based assay for which a (partial or complete) ladder of candidate solutions can be developed.

15 With gel electrophoresis, it is preferable to include on the gel at least one allelic ladder for every marker used. For example, one lane can be dedicated to ladders for the marker panel used in the other lanes; this lane can be loaded in duplicate on the gel. It is also desirable to include at least one known reference allele lane on every gel; this signal can be one or more 20 positive PCR controls. The advantage of duplicate control lanes (for ladders and reference controls) is that when there is a PCR, loading, detection, or other failure in one lane, the signal in the other lane can be used. Moreover, a comparison of the two (or more) signals can suggest when such a failure may have occurred.

25 With capillary electrophoresis, it is similarly preferable to use ladder and reference controls.

- With single capillary systems (such as the ABI/310), these controls should be run at some point during the lifetime of

the capillary. Preferably, the controls should be run as often as the temporal variation in the sizing system (i.e., differential sample and size standard migration) warrant.

- With a multiple capillary instrument (e.g., ABI/3700, 5 MegABACE, etc.) each capillary can form its own electrophoretic separation system. In the most preferred embodiment, the allelic ladder, known reference samples, and any other capillary controls (for fragment sizing, color separation, etc.) are run at least once for every capillary, 10 with the calibration results then applied specifically to that capillary. For example, consider the case of using one panel comprised of a set of markers, with this panel applied to a set of s samples, run out on an n -capillary instrument that achieves r sequential runs per capillary with acceptable sizing fidelity. Suppose (for high-throughput studies) that, approximately, $n \times r \leq s$. Then, one of the runs (e.g., run number 1 or $r/2$) should have allelic ladders in all n capillaries, and another run (e.g., run number 2 or $1+r/2$) should have allele references in all n capillaries. 20

0
10
20
30
40
50
60
70
80
90

Generating data containing these allelic ladder sizing controls, and analyzing the data as described in this step, reduce run-to-run sizing variation. Such reduction in gels or capillaries is crucial for achieving reproducible sizing assays. 25 Variable run conditions (e.g., temperature, gel consistency, formamide concentration, sample purity, concentration, capillary length, gel thickness, etc.) can induce differential sizing between the PCR products in the sample and the internal size standards. These allelic ladder comparison methods help correct 30 for such variation.

Referring to Figure 5, Step 8 is for transforming coordinates.

In the preferred embodiment, size comparisons used in the analysis are performed in the allelic ladder size coordinate system, rather than in the size standard coordinate system. While the latter approach is also workable, the former method has the advantages that the reference system is comprised of DNA bands size calibrated to actual integer DNA molecule lengths, rather than to arbitrary fractional molecular weight sizing units. Using integer DNA lengths (e.g., true base pairs) is closer to the physical reality, and can simplify the data analysis, logical inference, communication with the user, quality assurance, and error checking.

05 15
10 20
15 25
20 30
25 35
30 40
35 45
40 50
45 55
50 60
55 65
60 70
65 75
70 80
75 85
80 90
85 95
90 100

In one preferred embodiment, the signals are kept in size standard units, but comparisons are made in the allelic ladder frame of reference. For example, in comparing sample peaks with allelic ladder reference peaks, the sample peaks can first be interpolated into a domain based on the allelic ladder peak sizes, prior to comparing the sample peaks with any other reference sizes.

A ladder-based peak sizing method can establish a direct connection between observed peak sizes and actual DNA fragment lengths (possibly up to a constant shift). Transforming data sizes into DNA lengths overcomes size binning problems. For example, rounding number to the nearest integer in DNA length (i.e., ladder) coordinates permits the assigning of consistent labels to each peak; this consistency is not achieved when round fractional peak size estimates. Moreover, the peak's deviation from the integer ladder provides a quality measure for how consistently the peaks are sizing on a particular size separation run.

In the most preferred embodiment, the sample signals are transformed from size standard units into allelic ladder DNA length units. Step 6 provides the sized signal profile of a lane (or capillary):

5 $f \circ \mu^{-1}(x_{size})$.

Step 7 provides a characterized allelic ladder that matches the alleles of the ladder in size standard units with corresponding DNA molecule lengths in base pairs (or other suitable integer-spaced units). This pairing defines a coordinate transformation
10 for that ladder:

$$\lambda: x_{size} \rightarrow x_{length}.$$

Combining the function $f \circ \mu^{-1}$ together with the function λ (this can be done using a double interpolation) forms:

$$f \circ \mu^{-1} \circ \lambda^{-1}(x_{length}),$$

which represents the signal intensity f in terms of DNA length x_{length} . The relevant mathematics and computer implementations are detailed above in Step 6, extracting the profiles. For each marker, with n lanes or capillaries, and k colors, these transformed profiles can be preferably stored in an $n \times k$ data structure (e.g., in memory, or in a file).

This procedure provides a method for analyzing a nucleic acid sample. The steps included: (a) forming labeled DNA sample fragments from a nucleic acid sample; (b) size separating and
25 detecting said sample fragments to form a sample signal; (c) forming labeled DNA ladder fragments corresponding to molecular lengths; (d) size separating and detecting said ladder fragments to form a ladder signal; (e) transforming the sample signal into length coordinates using the ladder signal; which in turn permit
30 (f) analyzing the nucleic acid sample signal in length coordinates, as follows.

Referring to Figure 5, Step 9 is for quantitating signal peaks.

In addition to sizing DNA peaks, it is also useful to quantitate the relative amount of DNA present. To do this accurately requires taking account of band overlap. Few systems currently perform this band overlap analysis; those that do (e.g., Cybergeneitics' TrueAllele software), use a combinatorial approach that increases analysis time greatly as the number of adjacent peaks increases. This combinatorial cost can impede analyses of large allelic ladders or of differential display data. In the preferred embodiment, as described herein, the DNA quantitation step for resolving band overlap should computationally scale (e.g., at linear or small polynomial cost) with the number of bands analyzed.

In DNA length coordinates, as developed in Step 8, a peak has a natural shape that stems from band broadening as it migrates down the gel or capillary. (For accurate peak quantitation, DNA length coordinates are preferable to size standard coordinates, and size standard coordinates are preferable to pixel scan coordinates.) Centered at location x_k , this peak shape can be described as a Normal function

$$\phi_k^L(x) \propto e^{-\frac{1}{2}\left(\frac{x-x_k}{\sigma}\right)^2}$$

on the leading edge, and a Cauchy function

$$25 \quad \phi_k^R(x) \propto \frac{1}{1 + \left(\frac{x - x_k}{\alpha} \right)^2}$$

on the receding edge; the functions share the same height h at x_k . Band broadening implies that as x_k increases, the width parameters σ and α will increase, the height h will decrease, while the area remains constant. This is confirmed by fitting peak spread data

as a function of DNA length with a third order exponential function.

Using the changing band shapes as a set of basis
5 functions, write the trace f (in DNA length coordinates) as:

$$f(x) = \sum_{k=1}^n c_k \cdot \phi_k(x)$$

where basis function ϕ_k depends on $\langle x_k, \sigma_k, \alpha_k, h_k \rangle$, i.e., the center position x_k , Normal spread σ_k , Cauchy spread α_k , and center height h_k .

10 From the data $f(x)$, solving for the coefficients $\{c_k\}$ provides an estimate of the DNA concentration at each peak. Less efficient approaches (e.g., TrueAllele) currently solve this equation by a least-squares search of the $4n$ coefficients to the data f , with combinatorial computing time proportional to $\exp(4n)$. However, by exploiting the model functions ϕ_k , together with function space mathematics, this computing time can be reduced to roughly *linear cost*, proportional to n . See (F Riesz and B Sz.-Nagy, Functional Analysis. New York: Frederick Ungar Publishing Co., 1952), incorporated by reference.

Normalize the basis functions ϕ_k so that

$$\langle \phi_k(x), \phi_k(x) \rangle \equiv 1$$

and note that the band overlap coefficients $b_{k,j}$ can be numerically
25 estimated from the model functions as

$$\langle \phi_k(x), \phi_j(x) \rangle \equiv b_{k,j}$$

Then observe that with initial estimates of x_k (assuming one peak per bp size k), rewrite the inner product $\langle f, \phi_j \rangle$ as:

$$\langle f, \phi_j \rangle = \left(\sum_{k=1}^n c_k \cdot \phi_k \right) \cdot \phi_j = \sum_{k=1}^n c_k \cdot (\phi_k \cdot \phi_j) = \sum_{k=1}^n c_k \cdot b_{k,j}$$

Setting the data derived values $\{d_j = \langle f, \phi_j \rangle\}$, and using appropriate vector-matrix notation, yields the relation:

$$\mathbf{d} = \mathbf{c} \times \mathbf{B}$$

5 With sparse peak data that has little band overlap (e.g., tetranucleotide repeat data) $\mathbf{B} = \mathbf{I}_n$ (the identity matrix), and $\mathbf{c} \approx \mathbf{d}$ immediately yields the DNA concentrations at every peak k. More generally, \mathbf{B} is largely an identity matrix (i.e., primarily zeros off-diagonal), but has a few small near-diagonal 10 elements (the band overlaps) added in. Solve for \mathbf{c} from the observed data vector \mathbf{d} and the band overlap matrix \mathbf{B} using a very fast matrix inversion algorithm (e.g., SVD) that exploits the sparse nature of the local overlap coefficients.

15 In the most preferred embodiment, the overlap matrix \mathbf{B} is used to rapidly estimate the DNA concentrations $\{c_k\}$ at the peaks. This is done by computationally exploiting the functional analysis. Since

$$\mathbf{d} = \mathbf{c} \times \mathbf{B},$$

20 rewriting yields:

$$\mathbf{d} \times \mathbf{B}^{-1} = \mathbf{c} \times \mathbf{B} (\times \mathbf{B}^{-1}) = \mathbf{c}$$

or

$$\mathbf{c} = \mathbf{d} \times \mathbf{B}^{-1};$$

25 and so the DNA concentrations \mathbf{c} can be estimated immediately from the values $\{d_j = \langle f, \phi_j \rangle\}$ derived from data signal f once the ϕ_j and their overlap integrals $\mathbf{B} = \langle \phi_j, \phi_k \rangle$ are known.

In the preferred embodiment, estimate the ϕ_j basis 20 functions and their overlaps. Each basis function ϕ_k depends on

$\langle x_k, \sigma_k, \alpha_k, h_k \rangle$. To reduce this four dimensional search to one dimension, proceed as follows.

- x_k can be estimated from the peak center.
- h_k is not a factor, since each ϕ_k is normalized to integrate to 1.
- σ_k and α_k , are empirically observed to be in a fixed ratio to one another (at least in local neighborhoods), due to the relatively constant peak shapes.

Therefore, a computationally efficient approach is to perform a one dimensional search for σ_k and α_k , keeping their ratio fixed. Then, if so desired, perform a quick local two dimensional search in the solution neighborhood to refine the values of σ_k and α_k .

To find the peak width parameters σ_k and α_k , first fix a value of σ in a local neighborhood; in 1D search, set α to a fixed proportion of σ ; in 2D search, set the value of α as well. For each observed peak, set the center x_k of ϕ_k to the peak's center location. Following normalization, the basis functions $\{\phi_k\}$ are determined. Then, for j 's in the neighborhood, compute $d_j = \langle f, \phi_j \rangle$, or:

$$\mathbf{d} = \langle f, \phi \rangle$$

Determine \mathbf{B} by numerical (or closed form) integration of the basis functions ϕ_j and ϕ_k :

$$25 \quad \mathbf{B} = [\langle \phi_k, \phi_j \rangle]$$

Compute \mathbf{B}^{-1} by inverting \mathbf{B} , or:

$$\mathbf{B}^{-1} = [\langle \phi_k, \phi_j \rangle]^{-1}$$

Note that terms far from the diagonal are essentially zero. These local neighborhood calculations of \mathbf{d} and \mathbf{B}^{-1} are rapid, and are used in the inner loop of the following minimization procedure.

5 Minimize the difference between the observed data signal f and the expected model function:

$$f(x) - \sum_{k=1}^n c_k \cdot \phi_k(x - x_k)$$

by substituting the c_k computed from:

$$\mathbf{c} = \mathbf{d} \times \mathbf{B}^{-1}$$

10 Minimization (e.g., using an L_2 norm) of the expression:

$$\left\| f(x) - \sum_{k=1}^n (\mathbf{d} \times \mathbf{B}^{-1})_k \cdot \phi_k(x - x_k) \right\|_2$$

in the neighborhood as σ (and possibly α) vary, finds the best estimate of the widths of the basis function. These widths can be modeled locally, interpolated by fitting with a cubic exponential function, and then used to generate appropriate basis functions across the full range of sizes.

15 Application of these modeled basis functions to the data function f produces robust estimates of the DNA concentration c_k at each DNA peak k , since, after convergence,

$$\mathbf{c} = \mathbf{d} \times \mathbf{B}^{-1}.$$

20 In the current art, workers use the heights or areas of the data peaks. Because of the extensive peak shape modeling done here, it is preferable to use the heights (i.e., c_k) or areas (determined from c_k and the ϕ_k peak shape, e.g., by closed form evaluation) of the computationally modeled peaks.

25 The match-based quantitation described herein is very well suited to *highly multiplexed data*, such as SNPs, differential display, DNA arrays, and pooled samples. This is because the

inner product operations (and local band overlap corrections) can be applied accurately and in constant time, regardless of the number of data peaks in the signal.

5 It is useful to assess the quality of the individual peak quantitation results. This assessment can be done by comparing the modeled data function

$$\hat{f}(x) = \sum_{k=1}^n c_k \cdot \phi_k(x)$$

10 with the observed data signal f . A normalized L_2 deviation (computed, e.g., via a correlation) between expected and observed based on the minimization search function above can be used as the comparison measure.

Referring to Figure 5, Step 10 is for analyzing the data, preferably by calling the alleles.

20 In the preferred embodiment, allelic (or other) DNA ladder data is available, and the alleles can be called by matching sample peaks relative to the ladder peaks. This match operation is fast, reliable, very accurate, and accounts for inter-gel or inter-capillary variations. Depending on the application, a window (typically ± 0.5 bp) is set around a ladder peak of calibrated DNA length. When a sample peak (preferably in ladder coordinates) lies within this bp size window, it can be
25 reliably designated as having the length of that ladder peak. Zero size deviation between the centers of sample and ladder peaks indicates a perfect match; greater deviations indicate a less reliable match.

30 In an alternative preferred embodiment, the data profiles can be stored in size coordinates, and brought into length coordinates only when needed. This is done by retrieving a sample's size coordinate profile, as well as and the ladder peaks.

The sample's peaks are found in size coordinates, and then interpolated in into length coordinates using the ladder size to peak mapping, as described previously. The sample's peaks are then in length coordinates, and can be rounded to the nearest 5 integer, or matched against integers representing valid allele designations. The deviation from such integers (e.g., the fractional component) can be used a measure of quality, accuracy, or concern.

10 In other embodiments, found in the current art's fragment sizing software, the peak sizes (e.g., the alleles in genotyping applications) are analyzed in a size standard coordinate system, and ladder calibration data is not used by the computer. This analysis entails a bottom-up collection and comparison of data from many samples to form data-directed bins. These bin distributions are then used to designate sample peak size. However, the distribution variance across multiple electrophoresis runs on different samples can be quite high. When these bins have overlapping sizes, allele size designation becomes quite uncertain.

15 20

In the preferred embodiment, both DNA length and amount are used in scoring the data. With STR genotyping, for example, where there can be multiple peaks, the DNA concentration (e.g., 25 modeled peak height or area) is used as measure of confidence in the observed peak. With a heterozygote genotype, two peaks are expected, and with a homozygote, only one. Since other peaks may be present from mixtures, PCR artifacts, or other DNA sources, the analysis will focus on the higher concentration peaks, 30 particularly those peaks residing in allelic ladder windows. Most of the peak signal mass should be concentrated in the most confident peaks: high DNA amount, and in a ladder window. When this is not the case, confidence in the data is lower.

Once the peaks have been quantitated at known DNA lengths, the data can be further analyzed. Such analyses include stutter deconvolution, relative amplification correction, allele calling, allele frequency determination (from pooled samples), 5 differential display comparisons, and mutation detection. In genotyping applications, allele calling should be done on the signals only after corrections (e.g., for stutter or relative amplification) have been made. See (MW Perlin, "Method and system for genotyping," U.S. Patent #5,541,067, Jul. 30, 1996; MW Perlin, 10 "Method and system for genotyping," U.S. Patent #5,580,728, Dec. 3, 1996; S-K Ng, "Automating computational molecular genetics: solving the microsatellite genotyping problem," Carnegie Mellon University, Doctoral dissertation CMU-CS-98-105, January 23, 1998), incorporated by reference.

0
15
20
25
30

In one preferred embodiment, allele calling on quantitated corrected data is done by:

1. Finding the largest peak (area or height), and ensuring that is within a window on the allelic ladder.
2. Removing all peaks from the signal that either (a) have a DNA length that is not in a window of the allelic ladder, or (b) have a DNA amount that is not within some minimum percentage of the largest peak.
3. Calling the alleles by matching the DNA lengths of each sample peak to the DNA sizing windows on the allelic ladder.
4. Applying rules to check for possible data artifacts. Typical rules are described below.
5. Computing a quality score, particularly for those data apparently free of data artifacts. Various quality score components are discussed below.
6. Recording the designated alleles, and the quality of the result.

30

Referring to Figure 6, the results of ladder processing, peak quantitation and allele calling can be visualized.

There are many "junk-detecting" rules that can be designed

5 and applied to data. Critical to all such rules is the ability to compare observed measures against expected behaviors. By modeling the steps of the processing, computing appropriate quality scores at each step, and comparing these observed data features with normative results, the invention enables a precise computer

10 diagnosis of problems with data signals and their quality. Some example rules include:

- Noise only. Using measures (such as Wilcoxon's signed-rank statistic) to test for randomness, a computer program can decide that an experiment produced primarily noise (BW Lindgren, Statistical Theory, Fourth Edition. New York, NY: Chapman & Hall, 1993), incorporated by reference.
- Low signal. The peaks should have heights over a certain (user-defined) minimum threshold. When a profile's highest peak does not reach that threshold, flag the problem.
- High signal. The peaks should not be over a certain (user-defined) maximum threshold. When a profile's highest peak does exceeds that threshold, flag the problem.
- Peak dispersion. The designated peaks should comprise a certain percentage of the total signal. If a profile's designated peaks contain less than that percentage, fire the rule.
- Algorithm conflict. When multiple scoring algorithms are used, they should agree on the scoring results. Report any conflicts.
- Relative height. For some applications (e.g., forensic STR analysis) the relative peak heights should be within a predefined ratio of each other. Indicate when a genotype has a second peak with a relative height that is too low.

0
5
10
15
20
25
30

- Third peak. One (homozygote) or two peaks (heterozygote) should contain most of the DNA signal. Indicate when a genotype has a third peak that contains too much DNA signal.
- Off ladder. All the allele peaks should be close to their ladder peaks. When one (or more) of the alleles are too far away from their ladder peak, inform the user.
- Uncorrelated. When there are two allele peaks, their centers should deviate similarly from their respective ladder peaks. When a genotype has deviations that do not correlate (i.e., their difference exceeds some threshold), flag the problem.
- Control check. The calls computed for a control experiment should be consistent with the known results. When they are not, flag the result.

Note that some of these rules (e.g., "off-ladder") make use of allelic ladders, when they are present.

To better understand the decision support role of each numerical quality score (such as those used in rules), and their decision thresholds, it is useful to collect the scores for a large set of data and analyze them. Collection proceeds by recording the set of scores for each applicable genotype result, and indexing the genotypes (say, by sample, gel, and locus).

25 Consider each such numerical score as a mapping from the genotypes to \mathbb{R}^1 . Classify the genotypes by how they were scored; one preferred classification includes unscorable (e.g., noise), low quality (e.g., rule firings), correctly called good data (e.g., human agreement), and incorrectly called good data (e.g., human agreement). The result classification can be obtained by comparing the computer calls (and rule firings) with human edited (or otherwise independently scored) data. On some useful subset of genotypes (e.g., each locus examined separately), the numerical

30

quality scores can be collated and histogrammed for each result classification; this produces a set of distributions, one for each classification. Comparing (numerically, visually, statistically, etc.) the different distribution profiles for the different 5 classifications provides insight into the utility and scaling of a numerical quality score, permitting the derivation of decision thresholds or probability models. In a preferred embodiment, a decision threshold is statistically determined by distinguishing two score distributions (e.g., correct and incorrect results) 10 according to a determined sensitivity or specificity. In another preferred embodiment, linear or logistic regression is used to model the probability of an accurate allele call. These thresholds or probabilities can be displayed to a user for enhanced confidence or decision support.

It is useful to compute a quality score on the good data; one criterion for "good data" is that the experiment does not trigger any "junk" detecting rules. Quality scores enable the ranking of experiment results for selective review, as well as the computation of accuracy probabilities. Many possible quality scores can be developed, depending on the application and the available data. In all cases, there is some kind of comparison between expected behavior and an observed result - small 25 deviations indicate high quality, whereas large deviations suggest lower quality. Example quality scores include:

- Domain measures. When ladder data is present, it is possible to compute the deviation between a candidate allele peak center and its associated ladder peak center. When ladder data is not present, a similar comparison can be made relative to a precalibrated bin center, rather than a ladder peak center. One useful score is the maximum (over the scored alleles) of the absolute value of the deviations. When this number is close to zero, the 30

confidence is high. As it increases to 0.5bp, the result becomes less confident.

- Range measures. A sizing data result pertains to an particular number of peaks; any additional peaks represent a dispersion of the signal mass away from the hypothesized score. Ideally, all the signal mass should be present in the called peaks, which can be measured by the peak centering ratio = (called peaks)/(all peaks). When this ratio is near unity, the confidence is high. As it decreases to zero, the result becomes less confident.
- Product measures. A product of a domain measure with a range measure can be a sensitive indicator of quality. In one preferred embodiment, let a result scale with 1 as the highest quality, and 0 as the lowest quality. Rescale the domain measure above to map the score interval [0,0.5] to the quality interval [1,0], with all scores above 0.5 set to 0. Rescale the range measure above to map the score interval [0,1] to the quality interval [1,0]. Then form the product of these two rescaled scores, so that the result lies in the interval [0,1].
- Function measures. Once all data corrections have been made (e.g., for PCR stutter, relative amplification, peak shapes, peak sizes, peak center deviations, band overlap, etc.) on the fully quantitated and modeled signal, the inverse of these corrections can be applied to the size result to resynthesize the signal. Comparison of the resynthesized signal with the data signal provides a measure of how completely the analysis modeled the data - the residual deviation can measure lack of confidence in the result. One such comparison is a correlation between the synthesized and data signals. This correlation can be computed so that small values indicate confidence (e.g., using a normalized L_2 deviation), or so that larger values

indicate confidence (e.g., using a normalized inner product or statistical correlation measure).

The development of some quality scores has been described (MW Perlin, "Method and system for genotyping," U.S. Patent 5 #5,876,933, Mar. 2, 1999; S-K Ng, "Automating computational molecular genetics: solving the microsatellite genotyping problem," Carnegie Mellon University, Doctoral dissertation CMU-CS-98-105, January 23, 1998; B Pálsson, F Pálsson, M Perlin, H Gubjartsson, K Stefánsson, and J Gulcher, "Using quality measures to facilitate allele calling in high-throughput genotyping," Genome Research, vol. 9, no. 10, pp. 1002-1012, 1999), incorporated by reference.

Probabilities can be computed from quality scores. The individual components of the quality scores generally lie on a numerical scale. Histogramming and multivariate regression analysis can provide insight into the distribution of the correctly scored "success" data population relative to the distribution of the incorrectly scored "failure" data population along each of these measures. A logit transformation of this dichotomous outcome variable is useful, and provides [0,1] bounds for probability estimates, and the use of a binomial (rather than normal) distribution for the error analysis. By applying standard logistic regression, the key underlying independent variables can be elucidated. This logistic regression analysis can help determine the thresholds used in the "junk-detection" rules, and, for each experiment, can compute the probability of an accurate score from the observed variables. For example, the domain and range measures used above can be used as two independent variables, with the outcome being success or failure in the computer correctly calling the allele results. Logistic regression on these variables with a preanalyzed data set can be used to construct a correctness probability for allele calls on further data sets. See (A Agresti, Categorical Data Analysis. New

York, NY: John Wiley & Sons, 1990; DW Hosmer and S Lemeshow, Applied Logistic Regression. New York: John Wiley & Sons, 1989; Statistical Analysis Software, SAS Institute, Cary, NC; Statistical Package for the Social Sciences, SPSS Inc., Chicago, 5 IL), incorporated by reference.

Data Review

Once the automated computer scoring has completed, it is 10 often useful to have a person assess the results. Since reviewing perfectly scored data is an inessential step, the data review should optimally focus on reviewing the least confident (but scorably) data. The outcome of this focused review is typically a decision for each experiment's result: accept the result, modify the result, reject the result, plan to redo the experiment in the laboratory, and so on.

In order to focus the data review, it is helpful to have 20 a prioritization. In the preferred embodiment, the quality score or accuracy probability is used to rank the experiments. The review may arrange the suspect experiments in different subsets, e.g., grouped by marker, sample, equipment, personnel, time, location, etc. By reviewing the worst data first (i.e., rule firings, least quality scores), the user is enabled to focus on 25 evaluation and repair of only the suspect data. Not reviewing highly confident data frees up human operator time for other process tasks. Moreover, not reviewing unscorable data is similarly useful.

30 Using these methods, low quality data can be identified and classified. At the single datum level, individual results can be examined and better understood. At the data set level, problematic loci, samples or runs can be identified by examining

percentages of outcome indicators (e.g., rule firings, low quality scores, calling errors) relative to data stratifications. For example, the number of miscalls of known samples (arranged by locus and gel run) can identify problematic data trends early on.

5 This information can provide useful quality control feedback to the laboratory, which can help improve the overall quality of future data generation.

Referring to Figure 7, it is useful to present the results graphically. Visualizations of the experiments (e.g., gel lane tracking or color separation) and the signal traces help the user understand potential problems with the data, and possibilities for their correction (both for the particular experiment, as well as for the overall data generation process). In the preferred embodiment, multiple graphical visualizations for inspecting and reviewing genotype data include interfaces for:

- inspecting and annotating the raw gel data;
- inspecting and editing the automated lane tracking and sizing;
- assessing data quality and marker size windows;
- reviewing and editing automatically called alleles (by quality score priority, or other user-selected data rankings), preferably with all data and inferences visually presented in the context of the marker's allelic ladder;
- visually examining the data signal, preferably overlaid upon an allelic ladder;
- flexible examination of bleedthrough (or "pull-up") artifact;
- reviewing computed DNA quantitations and sizes, preferably in the context of the allelic ladder. A more detailed window explores the quantitation results.
- 25 • showing the allele calls;
- providing access to a more detailed textual window presenting a summary of useful allele calling information in tabular form;
- 30

- focusing on the allelic ladder data signals, when available; and
- displaying selected multiple lane signals graphically in one view.

5 Such multiple representations visually show the quality of the data, and provide diverse, focused insights into the data and their processing. Such graphical interfaces are generally used on only a small fraction (e.g., less than 5%-10%) of the data, since most high-throughput users do not care to revisit high-quality

10 allele calls on high-quality data.

Referring to Figure 8, it is also useful to present the results textually. Textual information can facilitate a more detailed understanding of a particular experiment. It is helpful for the user to have rapid access to both graphical and text presentation formats. The preferred embodiment provides information on allele designations, ladder comparisons, molecular weight sizes, and genotype quality. This display also gives explanations of rules that may have fired. The table includes peak size, peak height, peak area, and peak fit quality. The display is extensible, permitting modifications to the display that can draw from the expected and observed values that are computed for each scored genotype.

25 The preferred embodiment provides very flexible navigation between the graphical and textual view. A gel display permits viewing of any combination of gel image, size standard peak centers, or lane/size tracking grid, as well as editing capability. An allele call display provides navigation facilities

30 (e.g., buttons and menus) for rapidly selecting samples and loci, including zoom, slider, overlay, and relational options. With a single click, the user can examine the peak and ladder sizes of any designated allele. The interface also automates many of the

mundane display aspects (e.g., signal resizing, quality prioritization), thereby enabling rapid user navigation.

For maximum flexibility, the preferred embodiment
5 reports results to computers and people in multiple ways.

Preferred modalities include:

- Providing a flexible format (e.g., tabbed text files, or SQL queries) for seamless interaction with database, spreadsheet, laboratory information management system (LIMS), text editing, and other computer software.
- Providing such a flexible format for input information.
- Providing such a flexible format for output results.
- Providing such a flexible format for logging, audit, and error messages.
- Recording rule firings and quality scores (along with allele calls) in result files (e.g., tabbed text) in such a flexible format. The rule firing representation is extensible and backwardly compatible, so that it is easy to add more rules over time as more cases are observed.
- When a window focused on a genotype, displaying the fired rules (if any), thereby setting the context for why particular low quality data were rejected.
- Listing or explaining the fired rules in plain English in a textual window for the human operator.
- Listing the designations, allelic ladder information, molecular weight deviations (between the designation and its allelic ladder), quality score, or a table of peak sizes, areas, heights, and qualities.

30 Different data artifacts are typically associated with their own specific visualizations. For example, a signal containing size-designated standard peaks do not overlay properly on a different signals' size-designated standard peaks; referring to Figure 9, this improper overlay can be visualized by

superimposing the signals in different colors. For most data artifacts, human data reviewers painstakingly (a) construct an appropriate visual representation (e.g., an overlay of specific signals) to (b) confirm or disconfirm the presence of the

5 artifact. The more efficient approach of the preferred embodiment is to reverse this order. First (b), have the computer automatically determine (from rules or other quality scores) what data artifacts are present, and then (a) automatically construct visualizations that are customized to each specific artifact.

10 These data-directed visualizations can be opened up automatically, or displayed upon request.

Referring to Figures 4, 6, and 9, to efficiently develop a library of such data artifact customized displays, the invention includes a graphical language interpreter. An element of the display library is a message template that operationally characterizes how to display the artifact. This template is filled in by applying it to specific data. A message for the interpreter includes a set of n-tuples that describe the type of display, the source of data, a size range, fluorescent dye, or other useful display information. The data source preferably refers to the $n \times k$ representation of data signals. This interpreter then dispatches on the display type (e.g., data signal, vertical line, fitted curve, annotation, etc.) and

25 possible subtype (e.g., main data, size ladder, allelic ladder, known control, negative control, etc.) of each tuple to supply additional display information (e.g., drawing color, line style, line thickness, etc.) that is specific to that display type. Execution of the set of messages constructs and presents the

30 customized display to the user for its corresponding data artifact.

Analysis System

Referring to Figure 10, the present invention pertains to a system for analyzing a nucleic acid sample 102, as specified above. The system comprises a means 104 for forming labeled DNA sample fragments 106 from a nucleic acid sample. The system further comprises means 108 for size separating and detecting said sample fragments to form a sample signal 110, said separating and detecting means in communication with the sample fragments. The system further comprises means 112 for forming labeled DNA ladder fragments 114 corresponding to molecular lengths. The system further comprises means 116 for size separating and detecting said ladder fragments to form a ladder signal 118, said separating and detecting means in communication with the ladder fragments. The system further comprises means 120 for transforming the sample signal into length coordinates 122 using the ladder signal, said transforming means in communication with the signals. The system further comprises means 124 for analyzing the nucleic acid sample signal in length coordinates, said analyzing means in communication with the transforming means.

DRAFT
15
20

Special Applications

For many applications, it is useful to generate sizing results that are comparable across different DNA sequencer instruments. This *platform interoperability* is essential, for example, when creating a reference DNA database (e.g., for human identification in forensics, or multi-laboratory genetic analyses) with data comparisons from multiple laboratories. Should different instruments at different laboratories yield different sizing results for PCR products relative to the size standards, then such DNA reference resources become almost useless. The present invention uses sizing ladders based on such sample

fragments (e.g., allelic ladders, with or without known reference samples) in order to assure that sizing results are based on sample fragment sizes. Specifically, said sizing results are not based solely on electrophoretic migration of DNA fragments
5 relative to size standard molecules; such relative migration can be highly variable across different instruments, and even on the same instrument when different run conditions are employed. The present invention overcomes this limitation in the current art, uses novel computer-based scoring of properly calibrated data to
10 provide automated sizing of DNA fragments, and enables true interoperability between different sizing instruments.

50
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
105
110
115
120
125
130
135
140
145
150
155
160
165
170
175
180
185
190
195
200
205
210
215
220
225
230
235
240
245
250
255
260
265
270
275
280
285
290
295
300
305
310
315
320
325
330
335
340
345
350
355
360
365
370
375
380
385
390
395
400
405
410
415
420
425
430
435
440
445
450
455
460
465
470
475
480
485
490
495
500
505
510
515
520
525
530
535
540
545
550
555
560
565
570
575
580
585
590
595
600
605
610
615
620
625
630
635
640
645
650
655
660
665
670
675
680
685
690
695
700
705
710
715
720
725
730
735
740
745
750
755
760
765
770
775
780
785
790
795
800
805
810
815
820
825
830
835
840
845
850
855
860
865
870
875
880
885
890
895
900
905
910
915
920
925
930
935
940
945
950
955
960
965
970
975
980
985
990
995
1000
1005
1010
1015
1020
1025
1030
1035
1040
1045
1050
1055
1060
1065
1070
1075
1080
1085
1090
1095
1100
1105
1110
1115
1120
1125
1130
1135
1140
1145
1150
1155
1160
1165
1170
1175
1180
1185
1190
1195
1200
1205
1210
1215
1220
1225
1230
1235
1240
1245
1250
1255
1260
1265
1270
1275
1280
1285
1290
1295
1300
1305
1310
1315
1320
1325
1330
1335
1340
1345
1350
1355
1360
1365
1370
1375
1380
1385
1390
1395
1400
1405
1410
1415
1420
1425
1430
1435
1440
1445
1450
1455
1460
1465
1470
1475
1480
1485
1490
1495
1500
1505
1510
1515
1520
1525
1530
1535
1540
1545
1550
1555
1560
1565
1570
1575
1580
1585
1590
1595
1600
1605
1610
1615
1620
1625
1630
1635
1640
1645
1650
1655
1660
1665
1670
1675
1680
1685
1690
1695
1700
1705
1710
1715
1720
1725
1730
1735
1740
1745
1750
1755
1760
1765
1770
1775
1780
1785
1790
1795
1800
1805
1810
1815
1820
1825
1830
1835
1840
1845
1850
1855
1860
1865
1870
1875
1880
1885
1890
1895
1900
1905
1910
1915
1920
1925
1930
1935
1940
1945
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
2025
2030
2035
2040
2045
2050
2055
2060
2065
2070
2075
2080
2085
2090
2095
2100
2105
2110
2115
2120
2125
2130
2135
2140
2145
2150
2155
2160
2165
2170
2175
2180
2185
2190
2195
2200
2205
2210
2215
2220
2225
2230
2235
2240
2245
2250
2255
2260
2265
2270
2275
2280
2285
2290
2295
2300
2305
2310
2315
2320
2325
2330
2335
2340
2345
2350
2355
2360
2365
2370
2375
2380
2385
2390
2395
2400
2405
2410
2415
2420
2425
2430
2435
2440
2445
2450
2455
2460
2465
2470
2475
2480
2485
2490
2495
2500
2505
2510
2515
2520
2525
2530
2535
2540
2545
2550
2555
2560
2565
2570
2575
2580
2585
2590
2595
2600
2605
2610
2615
2620
2625
2630
2635
2640
2645
2650
2655
2660
2665
2670
2675
2680
2685
2690
2695
2700
2705
2710
2715
2720
2725
2730
2735
2740
2745
2750
2755
2760
2765
2770
2775
2780
2785
2790
2795
2800
2805
2810
2815
2820
2825
2830
2835
2840
2845
2850
2855
2860
2865
2870
2875
2880
2885
2890
2895
2900
2905
2910
2915
2920
2925
2930
2935
2940
2945
2950
2955
2960
2965
2970
2975
2980
2985
2990
2995
3000
3005
3010
3015
3020
3025
3030
3035
3040
3045
3050
3055
3060
3065
3070
3075
3080
3085
3090
3095
3100
3105
3110
3115
3120
3125
3130
3135
3140
3145
3150
3155
3160
3165
3170
3175
3180
3185
3190
3195
3200
3205
3210
3215
3220
3225
3230
3235
3240
3245
3250
3255
3260
3265
3270
3275
3280
3285
3290
3295
3300
3305
3310
3315
3320
3325
3330
3335
3340
3345
3350
3355
3360
3365
3370
3375
3380
3385
3390
3395
3400
3405
3410
3415
3420
3425
3430
3435
3440
3445
3450
3455
3460
3465
3470
3475
3480
3485
3490
3495
3500
3505
3510
3515
3520
3525
3530
3535
3540
3545
3550
3555
3560
3565
3570
3575
3580
3585
3590
3595
3600
3605
3610
3615
3620
3625
3630
3635
3640
3645
3650
3655
3660
3665
3670
3675
3680
3685
3690
3695
3700
3705
3710
3715
3720
3725
3730
3735
3740
3745
3750
3755
3760
3765
3770
3775
3780
3785
3790
3795
3800
3805
3810
3815
3820
3825
3830
3835
3840
3845
3850
3855
3860
3865
3870
3875
3880
3885
3890
3895
3900
3905
3910
3915
3920
3925
3930
3935
3940
3945
3950
3955
3960
3965
3970
3975
3980
3985
3990
3995
4000
4005
4010
4015
4020
4025
4030
4035
4040
4045
4050
4055
4060
4065
4070
4075
4080
4085
4090
4095
4100
4105
4110
4115
4120
4125
4130
4135
4140
4145
4150
4155
4160
4165
4170
4175
4180
4185
4190
4195
4200
4205
4210
4215
4220
4225
4230
4235
4240
4245
4250
4255
4260
4265
4270
4275
4280
4285
4290
4295
4300
4305
4310
4315
4320
4325
4330
4335
4340
4345
4350
4355
4360
4365
4370
4375
4380
4385
4390
4395
4400
4405
4410
4415
4420
4425
4430
4435
4440
4445
4450
4455
4460
4465
4470
4475
4480
4485
4490
4495
4500
4505
4510
4515
4520
4525
4530
4535
4540
4545
4550
4555
4560
4565
4570
4575
4580
4585
4590
4595
4600
4605
4610
4615
4620
4625
4630
4635
4640
4645
4650
4655
4660
4665
4670
4675
4680
4685
4690
4695
4700
4705
4710
4715
4720
4725
4730
4735
4740
4745
4750
4755
4760
4765
4770
4775
4780
4785
4790
4795
4800
4805
4810
4815
4820
4825
4830
4835
4840
4845
4850
4855
4860
4865
4870
4875
4880
4885
4890
4895
4900
4905
4910
4915
4920
4925
4930
4935
4940
4945
4950
4955
4960
4965
4970
4975
4980
4985
4990
4995
5000
5005
5010
5015
5020
5025
5030
5035
5040
5045
5050
5055
5060
5065
5070
5075
5080
5085
5090
5095
5100
5105
5110
5115
5120
5125
5130
5135
5140
5145
5150
5155
5160
5165
5170
5175
5180
5185
5190
5195
5200
5205
5210
5215
5220
5225
5230
5235
5240
5245
5250
5255
5260
5265
5270
5275
5280
5285
5290
5295
5300
5305
5310
5315
5320
5325
5330
5335
5340
5345
5350
5355
5360
5365
5370
5375
5380
5385
5390
5395
5400
5405
5410
5415
5420
5425
5430
5435
5440
5445
5450
5455
5460
5465
5470
5475
5480
5485
5490
5495
5500
5505
5510
5515
5520
5525
5530
5535
5540
5545
5550
5555
5560
5565
5570
5575
5580
5585
5590
5595
5600
5605
5610
5615
5620
5625
5630
5635
5640
5645
5650
5655
5660
5665
5670
5675
5680
5685
5690
5695
5700
5705
5710
5715
5720
5725
5730
5735
5740
5745
5750
5755
5760
5765
5770
5775
5780
5785
5790
5795
5800
5805
5810
5815
5820
5825
5830
5835
5840
5845
5850
5855
5860
5865
5870
5875
5880
5885
5890
5895
5900
5905
5910
5915
5920
5925
5930
5935
5940
5945
5950
5955
5960
5965
5970
5975
5980
5985
5990
5995
6000
6005
6010
6015
6020
6025
6030
6035
6040
6045
6050
6055
6060
6065
6070
6075
6080
6085
6090
6095
6100
6105
6110
6115
6120
6125
6130
6135
6140
6145
6150
6155
6160
6165
6170
6175
6180
6185
6190
6195
6200
6205
6210
6215
6220
6225
6230
6235
6240
6245
6250
6255
6260
6265
6270
6275
6280
6285
6290
6295
6300
6305
6310
6315
6320
6325
6330
6335
6340
6345
6350
6355
6360
6365
6370
6375
6380
6385
6390
6395
6400
6405
6410
6415
6420
6425
6430
6435
6440
6445
6450
6455
6460
6465
6470
6475
6480
6485
6490
6495
6500
6505
6510
6515
6520
6525
6530
6535
6540
6545
6550
6555
6560
6565
6570
6575
6580
6585
6590
6595
6600
6605
6610
6615
6620
6625
6630
6635
6640
6645
6650
6655
6660
6665
6670
6675
6680
6685
6690
6695
6700
6705
6710
6715
6720
6725
6730
6735
6740
6745
6750
6755
6760
6765
6770
6775
6780
6785
6790
6795
6800
6805
6810
6815
6820
6825
6830
6835
6840
6845
6850
6855
6860
6865
6870
6875
6880
6885
6890
6895
6900
6905
6910
6915
6920
6925
6930
6935
6940
6945
6950
6955
6960
6965
6970
6975
6980
6985
6990
6995
7000
7005
7010
7015
7020
7025
7030
7035
7040
7045
7050
7055
7060
7065
7070
7075
7080
7085
7090
7095
7100
7105
7110
7115
7120
7125
7130
7135
7140
7145
7150
7155
7160
7165
7170
7175
7180
7185
7190
7195
7200
7205
7210
7215
7220
7225
7230
7235
7240
7245
7250
7255
7260
7265
7270
7275
7280
7285
7290
7295
7300
7305
7310
7315
7320
7325
7330
7335
7340
7345
7350
7355
7360
7365
7370
7375
7380
7385
7390
7395
7400
7405
7410
7415
7420
7425
7430
7435
7440
7445
7450
7455
7460
7465
7470
7475
7480
7485
7490
7495
7500
7505
7510
7515
7520
7525
7530
7535
7540
7545
7550
7555
7560
7565
7570
7575
7580
7585
7590
7595
7600
7605
7610
7615
7620
7625
7630
7635
7640
7645
7650
7655
7660
7665
7670
7675
7680
7685
7690
7695
7700
7705
7710
7715
7720
7725
7730
7735
7740
7745
7750
7755
7760
7765
7770
7775
7780
7785
7790
7795
7800
7805
7810
7815
7820
7825
7830
7835
7840
7845
7850
7855
7860
7865
7870
7875
7880
7885
7890
7895
7900
7905
7910
7915
7920
7925
7930
7935
7940
7945
7950
7955
7960
7965
7970
7975
7980
7985
7990
7995
8000
8005
8010
8015
8020
8025
8030
8035
8040
8045
8050
8055
8060
8065
8070
8075
8080
8085
8090
8095
8100
8105
8110
8115
8120
8125
8130
8135
8140
8145
8150
8155
8160
8165
8170
8175
8180
8185
8190
8195
8200
8205
8210
8215
8220
8225
8230
8235
8240
8245
8250
8255
8260
8265
8270
8275
8280
8285
8290
8295
8300
8305
8310
8315
8320
8325
8330
8335
8340
8345
8350
8355
8360
8365
8370
8375
8380
8385
8390
8395
8400
8405
8410
8415
8420
8425
8430
8435
8440
8445
8450
8455
8460
8465
8470
8475
8480
8485
8490
8495
8500
8505
8510
8515
8520
8525
8530
8535
8540
8545
8550
8555
8560
8565
8570
8575
8580
8585
8590
8595
8600
8605
8610
8615
8620
8625
8630
8635
8640
8645
8650
8655
8660
8665
8670
8675
8680
8685
8690
8695
8700
8705
8710
8715
8720
8725
8730
8735
8740
8745
8750
8755
8760
8765
8770
8775
8780
8785
8790
8795
8800
8805
8810
8815
8820
8825
8830
8835
8840
8845
8850
8855
8860
8865
8870
8875
8880
8885
8890
8895
8900
8905
8910
8915
8920
8925
8930
8935
8940
8945
8950
8955
8960
8965
8970
8975
8980
8985
8990
8995
9000
9005
9010
9015
9020
9025
9030
9035
9040
9045
9050
9055
9060
9065
9070
9075
9080
9085
9090
9095
9100
9105
9110
9115
9120
9125
9130
9135
9140
9145
9150
9155
9160
9165
9170
9175
9180
9185
9190
9195
9200
9205
9210
9215
9220
9225
9230
9235
9240
9245
9250
9255
9260
9265
9270
9275
9280
9285
9290
9295
9300
9305
9310
9315
9320
9325
9330
9335
9340
9345
9350
9355
9360
9365
9370
9375
9380
9385
9390
9395
9400
9405
9410
9415
9420
9425
9430
9435
9440
9445
9450
9455
9460
9465
9470
9475
9480
9485
9490
9495

to variable size standard comigration units), it is possible to reliably match genotypes of a given sample with those of a reference sample. This comparison and reporting is enabled by the present invention. Moreover, this reliability is essential for 5 human identification applications where near zero error is required. In criminal comparisons, the sample DNA profile may be from a stain at a crime scene, whereas the reference DNA profile is from a suspect or a preexisting DNA database, with the goal of establishing a connection between a suspect and a crime scene. In 10 civil applications, the sample and reference DNA profiles may be used to determine degree of relatedness, as in a paternity case. (IW Evett and BS Weir, *Interpreting DNA Evidence : Statistical Genetics for Forensic Scientists*. Sunderland, MA: Sinauer Assoc, 1998), incorporated by reference.

15
16
17
18
19
20
21
22
23
24
25

PCR artifacts can complicate DNA fragment sizing. With STRs, PCR stutter, relative amplification (also termed "preferential amplification"), and constant bands are common artifacts. Using computer-based scoring methods, these artifacts can be resolved by stutter deconvolution, adjusting relative peak heights based on fragment size, detecting and suppressing artifactual bands, and other quantitative methods. (MW Perlin, "Method and system for genotyping," U.S. Patent #5,541,067, Jul. 30, 1996; MW Perlin, "Method and system for genotyping," U.S. 25 Patent #5,580,728, Dec. 3, 1996; S-K Ng, "Automating computational molecular genetics: solving the microsatellite genotyping problem," Carnegie Mellon University, Doctoral dissertation CMU-CS-98-105, January 23, 1998), incorporated by reference.

30 *Differential display* is a sensitive assay for relative gene expression. In automated computer data analysis of differential display gene expression profiles, the objective is to identify size bins at which there is a demonstrable difference

between the DNA amounts present in the two profiles. In a preferred embodiment, to compare:

- Transform the expression profiles so that each resides in a uniform DNA sizing coordinate system, referring to Figure 1, Step 6. When the pixel representation is highly reproducible (e.g., when running serially on a single capillary), this transforming step can be omitted.
- Identify the peaks in each profile, and record their x domain (DNA size) and y height (estimated DNA concentration) values.
- Compare the domain values of the peaks in each profile, and form a set of matching paired peaks (one from each profile). Retain those paired peaks with close x values. In the preferred embodiment, the difference between the x values is less than or equal to 1/2 base pair.
- Retain the cross-profile peak pairs having relatively close y value ratios. Compute the standard deviation of the ratio of the y values for peak pairs, and remove those peak pairs whose y-value ratios lie outside a certain standard deviation range (e.g., one standard deviation).
- Rescale the profile ranges so that they approximately superimpose. Normalize the first profile by dividing by its maximum y value. Model the y-value ratios polynomially (e.g., linearly) as a function of x. Normalize the second profile by dividing by the modeled y-value ratio function.
- Refine the peak matching by a two dimensional comparison that requires the matching peak pairs to satisfy both a certain x-tolerance (e.g., 0.5 bp) and certain y tolerance (e.g., 5% relative height). When the 1D peak matching does not produce spurious results, this refining step may be omitted.
- Use the matched peaks as a peak ladder, and transform the coordinates of one profile into the coordinates of the second profile, referring to Figure 1, Step 8. When the

peaks are very close (preferably less than 0.25 bp deviation), this transforming step may be omitted.

- Compare the superimposed profiles, and identify peak pairs whose calculated y-value differences or ratios are outliers, relative to the paired y value standard deviations. These outlying peak pairs represent possibly significant up- or down-regulation of gene expression.
- Report the identified outlying peak pairs. This can be done as two lists (one for up-regulation, and one for down-regulation). Within each list, rank the results by the y value deviation.

The results of such an analysis are shown in Figure 11, which demonstrates no evident variation in the repeated running of one differential display sample.

Software Description

Referring to Figure 12, the data scoring software is preferably maintained in a version control system. After testing has completed, program changes are committed. For each supported platform (Macintosh, Windows, Unix, etc.), an automated assembly computer retrieves the updated software and supporting data from the version control server, preferably over a computer network. Then, this computer compiles the software and data for run time operation on the specific target platform, and follows automation scripts to assemble the materials into an installer process, preferably for CD-ROMs or internet installation (e.g., MindVision's VISE for the Macintosh, or InstallShield for Windows). Hypertext documentation for the software is maintained, updated, and compiled in a cross-platform format, such as bookmarked PDF files, preferably using automated document authoring software (e.g., Adobe's FrameMaker program). For each supported platform, the application program, associated data, and

documentation are included on the web or disk installers. Users preferably download or update their software using the web internet installer; alternatively, a disk installer can be used locally on a computer or local area network.

5

The automated scoring software maintains an audit trail of its actions, operation, and decisions as it processes the data according the steps of Figures 1 and 5. The data formats are kept operational for specific platforms by automatically checking and 10 transforming certain files (e.g., text representations) prior to the program's using these files on that platform. The data format permits multiple instrument data acquisition runs (e.g., gel or capillary loadings) to be processed together in a single computer analysis run. These software features reduce human intervention 15 and error.

User feedback on the program's operation is preferably entered at a designated web site onto a reporting HTML form that includes expected and observed program behavior. Via CGI scripts, 20 these reports are automatically logged onto a database (e.g., FileMaker Pro on a Macintosh over a local area network), and appropriate personnel may be automatically notified via email of potential problems.

25 Business Considerations

The automated quality maintenance system described herein for generating and analyzing DNA fragment data has a nonobvious business model. It is desirable for groups to generate 30 and analyze their own data, since the users of the data often have the greatest incentive to maintain data quality. Moreover, people involved in the data generation task require continual feedback from the data analysis results. By generating data that includes

proper calibration reference standards (e.g., internal size standards, allelic ladders, reference traces, etc.), high quality data can be automatically analyzed in ways that lower quality data cannot.

5

The cost of manual scoring of data is quite high. The preferred business model includes a spreadsheet that permits an end-user to calculate their labor costs. Referring to Figure 13, a prospective or current customer can enter parameters related to 10 the cost of labor, the data generation throughput, and how effectively the labor force analyzes the generated data. From these factors, the spreadsheet can calculate the total labor cost, as well as the labor cost per experiment, for the specific customer requirements. This calculating tool is preferably made available to customers in spreadsheet form (e.g., as a platform-independent Microsoft Excel 98 spreadsheet) as a computer file that can be distributed on disk, via email attachment, or downloadable from the internet. In a complementary embodiment, the spreadsheet functionality can be provided as an interative 20 form on a web page. By better understanding the labor costs of data scoring, a customer can develop insight into the role of quality data and automated computer-based data scoring as a direct replacement for labor.

25 Error is another significant cost in the human scoring of genetic data. In genetics, error severely compromises the power of linkage and other association measures, so that despite considerable research time, cost, and effort, genes are less likely to be discovered. In forensics, error can lead to 30 incorrect identification of suspects, failed convictions of criminals, or failure in exonerating the innocent. Thus, methods that reduce error or improve data quality confer significant advantages to the user.

Since automated computer-based scoring of DNA sizing data is equivalent to human labor that performs the same task, the pricing model of such automated scoring is preferably based on usage. A fee is charged for every genotype scored. This fee 5 preferably includes components for intellectual property, software support, and user customizations. With very high levels of usage, some components such as user customization can be reduced in line with the associated business expense reduction.

10 For better market penetration, pricing levels should be set near or below the equivalent human labor cost. The result is an automated computer-based scoring process that is faster, better and cheaper than the equivalent human review of data. Specifically, the computer-based process can produce more 15 consistent results with lower error, leading to more productive use of the scored data.

20 Additional services (including user training, system setup, software modifications, quality audits) are best charged separately. Preferably, there is a mandatory initial interaction with the customer group (for which the user pays) to train new 25 users in the quality data generation and computer-based analysis process.

30 Quality assurance is an integral part of the process and the business model. A quality entity (e.g., a corporation) can help support this quality maintenance system (QMS). This entity can provide quality assurance by spot checking the user's data generation or scoring of the DNA fragment size data. This quality assurance can be done by (a) the entity providing the user with samples (characterized by the entity) for data generation or analysis, and then comparing the user's results with the entity's results, (b) the user providing the entity with samples (characterized by the user) for data generation or analysis, and

then having the entity comparing the entity's results with the user's results, (c) a comparison of data results involving a third party, or (d) a double-blind comparison study.

5 The quality entity can use these quality assurance methods to conduct quality audits for different sites, and disseminate the results (e.g., by internet publication). Then, different data generation sites can compete with one another for business based on their quality and cost-effectiveness. Beyond a
10 certain critical mass, it would be highly desirable for such DNA analysis sites to have certification provided by the quality entity on the basis of such audits. Those sites with the highest quality data and using the best automation software should be the most competitive. The quality entity can also provide a service (e.g., at regular time intervals, say annually) for quality checks on sites that desire to achieve and maintain the best possible data results.

20 This procedure describes a method for generating revenue from computer scoring of genetic data. The steps include: (a) supplying a software program that automatically scores genetic data; (b) forming genetic data that can be scored by the software program; (c) scoring the genetic data using the software program to form a quantity of genetic data; and (d) generating a revenue
25 from computer scoring of genetic data that is related to the quantity. Moreover, additional process steps include: (e) defining a labor cost of scoring the quantity of genetic data when not using the software program; (f) providing a calculating mechanism for estimating the labor cost from the quantity; (g) determining the labor cost based on the quantity; and (h) establishing a price for using the software program that is related to the labor cost.

Mixture Analysis

In forensic science, DNA samples are often derived from more than one individual. With the advent of quantitative analysis of STR data, there is the possibility of computer-based analysis that can resolve these data. Specifically, there is a need to find or confirm the identity of component DNA profiles, as well as determine mixture ratios. In the preferred embodiment, the quantitation of the DNA samples is accomplished by performing Steps 1-6 of Figure 1, and Steps 7-9 of Figure 5. The accurate quantitation conducted in Step 9 enables accurate analysis of quantitative mixture data, and improves on the prior art that uses unmodeled peak area or peak height (GeneScan Software, PE Biosystems, Foster City, CA) which provide potentially inaccurate estimates of DNA concentration. The invention's quantitative mixture analysis is part of Step 10 of Figure 5 in the case of DNA mixtures.

The present invention is distinguished over the prior art in that it uses a linear mathematical problem solving formulation that combines information across loci, and can completely integrate this information automatically on a computing device. By inherently using all the information from all the loci in the formulation, robust solutions can be achieved. The prior art uses manual or Bayesian methods on a per locus basis that entail human intervention in generating or combining partial results. Such prior forensic mixture analysis methods have been described (P Gill, R Sparkes, R Pinchin, TM Clayton, JP Whitaker, and J Buckleton, "Interpreting simple STR mixtures using allele peak area," *Forensic Sci. Int.*, vol. 91, pp. 41-53, 1998; IW Evett, P Gill, and JA Lambert, "Taking account of peak areas when interpreting mixed DNA profiles," *J. Forensic Sci.*, vol. 43, pp. 62-69, 1997; TM Clayton, JP Whitaker, R Sparkes, and R Gill, "Analysis and interpretation of mixed forensic stains using DNA

STR profiling," *Forensic Sci. Int.*, vol. 91, pp. 55-70, 1998), incorporated by reference.

There are different formulations of the mixture problem.

5 With a mixture profile derived from two individuals, problems include:

- verifying the mixture and computing the mixture ratio, given the profiles of both individuals;
- determining the profile of one individual and the mixture ratio, given the profile of another individual; and
- determining the profiles of both individuals and the mixture ratio, given no other DNA profiles.

10 These problems can be similarly extended to a number of individuals greater than two. The STR mixture method described herein addresses all of these problem formulations.

20 In the PCR amplification of a mixture, the amount of each PCR product scales in rough proportion to relative weighting of each component DNA template. This holds true whether the PCRs are done separately, or combined in a multiplex reaction. Thus, if two DNA samples A and B are in a PCR mixture with relative concentrations weighted as w_A and w_B ($0 \leq w_A, w_B \leq 1$, $w_A + w_B = 1$), their corresponding signal peaks after detection will generally have peak quantitations (height or area) showing roughly the same 25 proportion. Therefore, by observing the relative peak proportions, one can estimate the DNA mixture weighting. Note that mixture weights and ratios are interchangeable, since the mixture weight $\frac{[A]}{[A]+[B]}$ is in one-to-one correspondence with the mixture ratio $\frac{[A]}{[B]}$.

30

To mathematically represent the linear effect of the DNA sample weights (w_A , w_B , w_C , ...), write the linear equation:

$$\mathbf{p} = \mathbf{G} \times \mathbf{w},$$

where \mathbf{p} is the pooled profile column vector, each column i of genotype matrix \mathbf{G} represents the alleles in the genotype of individual i (taking allele values 0, 1, 2, with a total of 2 alleles), and \mathbf{w} is the weight column vector that reflects the relative proportions of template DNA or PCR product. To illustrate this coupling of DNA mixture weights with predicted pool weights, if there are three individuals A, B, C represented in a mixture with weighting $w_A=0.5$, $w_B=0.25$, $w_C=0.25$, and at one locus the genotypes are:

A has allele 1 and allele 2,
B has allele 1 and allele 3, and
C has allele 2 and allele 3,

then combining the vectors via the linear equations:

$$\begin{bmatrix} \text{alleles} \\ \text{in} \\ \text{mixture} \end{bmatrix} = \begin{bmatrix} \text{alleles} \\ \text{of} \\ \text{A} \end{bmatrix} \begin{bmatrix} \text{alleles} \\ \text{of} \\ \text{B} \end{bmatrix} \begin{bmatrix} \text{alleles} \\ \text{of} \\ \text{C} \end{bmatrix} \times \begin{bmatrix} w_A \\ w_B \\ w_C \end{bmatrix}$$

and representing each allele as a position in a column vector, produces the linear relationship:

$$\begin{bmatrix} 0.75 \\ 0.75 \\ 0.50 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} 0.50 \\ 0.25 \\ 0.25 \end{bmatrix}$$

Note that the sum of alleles in each allele column vector is normalized to equal two, the number of alleles present.

With multiple loci, the weight vector \mathbf{w} is identical across all the loci, since that is the underlying mixture in the DNA template. This coupling of loci can be represented in the linear equations by extending the column vectors \mathbf{p} and \mathbf{G} with more allele information for additional loci. To illustrate this coupling of DNA mixture weights across loci, add a second locus to the three individuals above, where at locus two the genotypes are:

A has allele 1 and allele 2,

B has allele 2 and allele 3, and

C has allele 3 and allele 4,

then combining the vectors via the partitioning:

$$\begin{bmatrix} locus1 \\ mixture \\ alleles \\ \hline locus2 \\ mixture \\ alleles \end{bmatrix} = \begin{bmatrix} locus1 \\ A's \\ alleles \\ \hline locus2 \\ A's \\ alleles \end{bmatrix} \begin{bmatrix} locus1 \\ B's \\ alleles \\ \hline locus2 \\ B's \\ alleles \end{bmatrix} \begin{bmatrix} locus1 \\ C's \\ alleles \\ \hline locus2 \\ C's \\ alleles \end{bmatrix} \times \begin{bmatrix} wA \\ wB \\ wC \end{bmatrix}$$

5 and representing each allele as a position in a column vector, produces:

$$\begin{bmatrix} 0.75 \\ 0.75 \\ 0.50 \\ \hline 0.50 \\ 0.50 \\ 0.75 \\ 0.50 \\ 0.25 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ \hline 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} 0.50 \\ 0.25 \\ 0.25 \end{bmatrix}$$

With multiple loci, there is more data and greater confidence in estimates computed from the linear equations.

10

More precisely, write the vector/matrix equation $\mathbf{p} = \mathbf{G} \times \mathbf{w}$ for mixture coupling (of individual and loci) as coupled linear equations that include the relevant data:

$$p_{ik} = \sum_j g_{ijk} w_j,$$

15 where for locus i , individual j , and allele k

- p_{ik} is the pooled proportion in the observed mixture of locus i , allele k ;
- g_{ijk} is the genotype of individual j at locus i in allele k , taking values 0 (no contribution), 1 (heterozygote or

hemizygote contribution), or 2 (homozygote contribution), though with anomalous chromosomes other integer values are possible; and

- w_j is the weighting in the mixture of individual j .

5

Given partial information about equation $\mathbf{p} = \mathbf{G} \times \mathbf{w}$, other elements can be computed by solving the equation. Cases include:

- When \mathbf{G} and \mathbf{w} are both known, then the data profile \mathbf{p} can be predicted. This is useful in search algorithms.
- When \mathbf{G} and \mathbf{p} are both known, then the weights \mathbf{w} can be computed. This is useful in confirming a suspected mixture, and in search algorithms.
- When \mathbf{p} is known, inferences can be made about \mathbf{G} and \mathbf{w} , depending on the prior information available (such as partial knowledge of \mathbf{G}). This is useful in human identification applications.

10

The DNA mixture is resolved in different ways, depending on the case.

15

Assume throughout that the mixture profile data vector \mathbf{p} has been normalized for each locus. That is, for each locus, let NumAlleles be the number of alleles found in data for that locus (typically NumAlleles = 2, one for each chromosome). For each allele element of the quantitation data within the locus, multiply by NumAlleles, and divide by the sum (over the alleles) of all the quantitation values for that locus. Then, the sum of the normalized quantitation data is NumAlleles, as in the illustrative example above.

20

To resolve DNA mixtures, perform the steps: (a) obtain DNA profile data that include a mixed sample; (b) represent the data in a linear equation; (c) derive a solution from the linear equation; and (d) resolve the DNA mixture from the solution. This procedure is illustrated in the following cases.

First consider the case where all the genotypes \mathbf{G} and the mixture data \mathbf{p} are known, and the mixture weights \mathbf{w} need to be determined. This problem is resolved by solving the linear equations $\mathbf{p} = \mathbf{G} \times \mathbf{w}$ for \mathbf{w} using SVD or some other matrix inversion method. Specifically, \mathbf{w} can be estimated as:

$$\mathbf{w} = \mathbf{G} \setminus \mathbf{p}$$

using the SVD matrix division operation in MATLAB.

Next consider the case of two individuals A and B where one of the two genotypes (say, A) is known, the mixture weights \mathbf{w} are known, and the quantitative mixture data profile \mathbf{p} is available. Expand $\mathbf{p} = \mathbf{G} \times \mathbf{w}$ in this case as:

$$\mathbf{p} = w_A \cdot \mathbf{g_A} + w_B \cdot \mathbf{g_B}$$

where $\mathbf{g_A}$ and $\mathbf{g_B}$ are the genotype column vectors of individuals A and B, and w_A and $w_B = (1-w_A)$ are their mixture weights. Then, to resolve the genotype, rewrite this equation as

$$\mathbf{g_B} = (\mathbf{p} - w_A \cdot \mathbf{g_A}) / w_B$$

and solve for $\mathbf{g_B}$ by matrix operations. The computed $\mathbf{g_B}$ is the normalized difference of the mixture profile minus A's genotype. The accuracy of the solution increases with the number of loci used, and the quality of the quantitative data.

Next consider the case of making inferences about the genotype matrix \mathbf{G} starting from a mixture profile \mathbf{p} . This case has utility in forensic science. In one typical scenario, a stain from a crime scene may contain a DNA mixture from the victim and an unknown individual, the victim's DNA is available, and the investigator would like to connect the unknown individual's DNA profile with a candidate perpetrator. This scenario typically occurs in rape cases. The perpetrator may be a specific suspect, or the investigator may wish to check the unknown individual's DNA profile against a DNA database of possible candidates. If the

mixture weight w_A were known, then the genotype \mathbf{g}_B could be computed immediately from the matrix difference operation of the preceding paragraph.

5 Since w_A is not known, a workable approach is to search for the best w in the $[0,1]$ interval that satisfies additional constraints on the problem, set w_A equal to this best w , compute the genotype $\mathbf{g}(w_A)$ as a function of this optimized w_A value, and set $\mathbf{g}_B = \mathbf{g}(w_A)$. A suitable constraint is the prior knowledge of
10 the form that possible solution genotype vectors \mathbf{g} can take. It is known that solutions must have a valid genotype subvector at each locus (e.g., having alleles taking on values 0, 1 or 2, and summing to 2). This knowledge can be translated into a heuristic function of $\mathbf{g}(w)$ which evaluates each candidate genotype solution \mathbf{g} against this criterion.

15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95

In the preferred embodiment, the heuristic is a function of w , the profile \mathbf{p} , and the known genotype \mathbf{g}_A . Since \mathbf{p} and \mathbf{g}_A are fixed for any given problem, in this case the function depends only on the variable w . For any given w in $(0,1)$, compute $\mathbf{g}(w)$ as $(\mathbf{p} - w\mathbf{g}_A) / (1-w)$. Then, at each locus, compute and record the deviation $dev_{locus}(\mathbf{g}(w))$. The dev_{locus} function at one locus is defined as:

25

- Assume the genotype comprises one allele. Compute the deviation by finding the index of the largest peak, and forming a vector **oneallele** that has the value 2 at this index and is 0 elsewhere. Let dev_1 be the L_2 norm of the difference between $\mathbf{g}(w)$ and **oneallele**.
- Assume the genotype comprises two alleles. Compute the deviation by finding the index of the two largest peaks, and forming a vector **twoallele** that has the value 1 at each of these two indices and is 0 elsewhere. Let dev_2

be the L_2 norm of the difference between $\mathbf{g}(w)$ and **twoallele**.

- Return the the lesser of the two deviations as $\min(\text{dev1}, \text{dev2})$.

5 To compute $\text{dev}(\mathbf{g}(w))$, sum the component $\text{dev}_{\text{locus}}(\mathbf{g}(w))$ at each locus. That is, the heuristic function is the scalar value

$$\text{dev}(\mathbf{g}(w)) = \sum_{\text{loci}} \text{dev}_{\text{locus}}(\mathbf{g}(w))$$

Appropriately optimize (e.g., minimize) this function over w in $[0,1]$ to find wB , and determine **gB** as $\mathbf{g}(wB)$. In one alternative 10 preferred embodiment, the summation terms can be normalized to reflect alternative preferred weightings of the loci or alleles. In a different alternative preferred embodiment, various heuristic functions can be used that reflect other reasonable constraints on the genotype vectors, as in (P Gill, R Sparkes, R Pinchin, TM Clayton, JP Whitaker, and J Buckleton, "Interpreting simple STR mixtures using allele peak area," *Forensic Sci. Int.*, vol. 91, pp. 41-53, 1998), incorporated by reference.

Referring to Step 10 of Figure 5, further quality 20 assessment can be performed on the computed STR profile derived from the mixture analysis. As described, rule checking can identify potentially anomalous allele calls, particularly when peak quantities or sizes do not conform to expectations. Quality 25 measures can be computed on the genotypes, which can indicate problematic calls even when no rule has fired. One preferred quality score in mixture analysis is the deviation $\text{dev}(\mathbf{gB})$ of the computed genotype. Low deviations indicate a good result, whereas high scores suggest a poor result. These deviations are best interpreted when scaled relative to a set of calibration data 30 which have been classified as correct or incorrect. Of particular utility is the partitioning of the deviations by locus, using the locus deviation function $\text{dev}_{\text{locus}}(\mathbf{gB})$. When a locus has an unusually high deviation, it can be dropped from the profile, and

the resulting partial profile can be used for human identity matching.

The most preferred embodiment is demonstrated here on
5 data generated using the 10 STR locus SGMplus panel (PE
BioSystems, Foster City, CA), and size separated and detected on
an ABI/310 genetic analyzer sequencing instrument. A mixture
proportion of 30% sample A and 70% sample B was used. Referring
10 to Figure 14, a minimization search for weight w by computing
 $dev(g(w))$ gave a weighting of 29.73%, which is very close to the
true 30%. The L_2 deviation $dev(g(w))$ of the computed genotype from
the closest (and correct) feasible solution was 0.3199. The
15 computed genotype for all the loci is shown in the columns PROFILE
(exact) and GENO B (rounded) in the table below.

20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95

Data and results are shown in the table below, where
MIXTURE is the normalized peak quantitation data from the mixed
sample, GENO A is the known genotype of individual A, PROFILE is
the numerical genotype computed for determining B's genotype,
GENO B is the resulting integer genotype (and, in this case, the
known genotype as well), and DEVS are the square deviations of the
PROFILE from GENO B. Quality assessment of the computed PROFILE
shows uniform peaks that are consistent with a correct genotype.
Examination of the square deviation components for each allele
25 reveals no significant problem, and the largest within locus sum
of squares deviation was the small value 0.16 (for locus D21S11).

| <u>LOCUS-ALLEL</u> | <u>MIXTURE</u> | <u>GENO A</u> | <u>PROFILE</u> | <u>GENO B</u> | <u>DEVS</u> |
|--------------------|----------------|---------------|----------------|---------------|-------------|
| D3S1358-14 | 1.0365 | 1.0000 | 1.0520 | 1.0000 | 0.0027 |
| 30 D3S1358-15 | 0.9635 | 1.0000 | 0.9480 | 1.0000 | 0.0027 |
| vWA-17 | 1.4755 | 0 | 2.0999 | 2.0000 | 0.0100 |
| vWA-18 | 0.5245 | 2.0000 | -0.0999 | 0 | 0.0100 |
| D16S539-11 | 1.4452 | 0 | 2.0567 | 2.0000 | 0.0032 |
| D16S539-13 | 0.2889 | 1.0000 | -0.0120 | 0 | 0.0001 |

| | | | | | | |
|----|--------------|--------|--------|---------|--------|--------|
| | D16S539-14 | 0.2660 | 1.0000 | -0.0447 | 0 | 0.0020 |
| | D2S1338-16 | 0.3190 | 1.0000 | 0.0308 | 0 | 0.0010 |
| | D2S1338-18 | 0.6339 | 0 | 0.9021 | 1.0000 | 0.0096 |
| | D2S1338-20 | 0.3713 | 1.0000 | 0.1052 | 0 | 0.0111 |
| 5 | D2S1338-21 | 0.6758 | 0 | 0.9618 | 1.0000 | 0.0015 |
| | D8S1179-9 | 0.7279 | 0 | 1.0359 | 1.0000 | 0.0013 |
| | D8S1179-12 | 0.2749 | 1.0000 | -0.0320 | 0 | 0.0010 |
| | D8S1179-13 | 0.6813 | 0 | 0.9696 | 1.0000 | 0.0009 |
| | D8S1179-14 | 0.3160 | 1.0000 | 0.0265 | 0 | 0.0007 |
| | D21S11-27 | 0.2787 | 1.0000 | -0.0265 | 0 | 0.0007 |
| 10 | D21S11-29 | 0.7876 | 0 | 1.1208 | 1.0000 | 0.0146 |
| | D21S11-30 | 0.9337 | 1.0000 | 0.9057 | 1.0000 | 0.0089 |
| | D18S51-12 | 0.3443 | 1.0000 | 0.0669 | 0 | 0.0045 |
| | D18S51-13 | 0.6952 | 0 | 0.9894 | 1.0000 | 0.0001 |
| | D18S51-14 | 0.6755 | 0 | 0.9613 | 1.0000 | 0.0015 |
| | D18S51-17 | 0.2850 | 1.0000 | -0.0176 | 0 | 0.0003 |
| 15 | D19S433-12.2 | 0.6991 | 0 | 0.9949 | 1.0000 | 0.0000 |
| | D19S433-14 | 0.6060 | 2.0000 | 0.0161 | 0 | 0.0003 |
| | D19S433-15 | 0.6949 | 0 | 0.9890 | 1.0000 | 0.0001 |
| | TH01-6 | 0.3178 | 1.0000 | 0.0291 | 0 | 0.0008 |
| | TH01-7 | 1.0074 | 1.0000 | 1.0105 | 1.0000 | 0.0001 |
| | TH01-9 | 0.6749 | 0 | 0.9605 | 1.0000 | 0.0016 |
| 20 | FGA-19 | 1.0580 | 1.0000 | 1.0826 | 1.0000 | 0.0068 |
| | FGA-24 | 0.2830 | 1.0000 | -0.0203 | 0 | 0.0004 |
| | FGA-25.2 | 0.6589 | 0 | 0.9378 | 1.0000 | 0.0039 |
| | | | | | | |
| | | | | | | |
| | | | | | | |

In an alternative preferred embodiment, the heuristic function can be extended to account for the curvature of deviations (as a function of w), so that only local minima are considered and not boundary points. Genotype elimination constraints can be applied when there is extra knowledge, such when the mixture weight (hence proportion of sample B) is low and sample A's genotype can be excluded in certain cases. It is also

possible to provide for interactive human feedback, so that expert assessments can work together with the computing method.

When stutter peaks are a concern, use stutter

5 deconvolution to mathematically remove the stutter artifact from the quantitative signal (MW Perlin, G Lancia, and S-K Ng, "Toward fully automated genotyping: genotyping microsatellite markers by deconvolution," Am. J. Hum. Genet., vol. 57, no. 5, pp. 1199-1210, 1995), incorporated by reference. The prior forensic art uses

10 Bayesian approaches to account for stutter (P Gill, R Sparkes, and JS Buckleton, "Interpretation of simple mixtures when artefacts such as stutters are present - with special reference to multiplex STRs using by the Forensic Science Service," Forensic Sci. Int., vol. 95, pp. 213-224, 1998), incorporated by reference. However, direct stutter removal from the signal can be highly robust, since it is working directly at the level of the stutter artifact, prior to any mixture computation.

Q
5
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95

In the case when both genotypes are unknown, use additional search. Start from the linear equations $\mathbf{p} = \mathbf{G} \times \mathbf{w}$. As necessary, form feasible submatrices \mathbf{H} (locus) of \mathbf{G} , where each \mathbf{H} is an $K \times 2$ matrix, representing K alleles (rows) for 2 individuals (columns). Here, $\mathbf{H} = \{g_{ijk}\}$, where locus i is fixed, individual $j \in \{1, 2\}$, and allele $k \in \{1, 2, \dots, K\}$. For example, the six

25 feasible four allele ($K=4$) genotype pairs are represented by the six genotype matrices $\{\mathbf{H}_i \mid i = 1, 2, \dots, 6\}$:

$$\mathbf{H} = \begin{bmatrix} g_{11} & g_{21} \\ g_{12} & g_{22} \\ g_{13} & g_{23} \\ g_{14} & g_{24} \end{bmatrix} \in \left\{ \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \right\}$$

Matrix division (e.g., SVD) is performed using these **H** submatrices. The mixture profile data at each individual locus **p**(locus) is also used. Proceed as follows:

1. Normalize the mixture data **p** to sum to 2 at each locus.
- 5 2. Find the best two element weighting vector **w**. On the subset of most informative loci (typically those loci having four allele loci):

For each locus,

For every valid genotype submatrix **H** at that locus,
10 compute $w(\text{locus}, H) = p(\text{locus}) \setminus H$
normalize $w(\text{locus}, H)$ to sum to 1
set $dev(\text{locus}, H)$ to $\text{norm}\{p(\text{locus}) - H \times w(\text{locus}, H)'\}$

Set **w** = $w(\text{locus}, H)$ having the smallest $dev(\text{locus}, H)$.

3. Derive the genotype **H** of each locus as follows:

For every valid genotype matrix **H_i** at that locus,

compute $dev(H_i) = \text{norm}\{p(\text{locus}) - H_i \times w'\}$

Set **H**(locus) to that **H_i** having the smallest $dev(H_i)$.

4. Assess the quality of the genotype result **G** formed by combining the $\{H(\text{locus})\}$ at each locus. This is preferably done by:

- (a) using the matrix operation $w = G \setminus p$ to estimate a second mixture weight **w**, and comparing with the first **w** found in Step 2, or
- (b) by examining the computed deviations $dev(H(\text{locus}))$ found in Step 3.

25

Note that the operation " $w(\text{locus}, H) = p(\text{locus}) \setminus H$ " in this procedure for computing the best weight uses a matrix operation (i.e., matrix division). This procedure is preferable to finding **w** by minimizing $\text{norm}\{p(\text{locus}) - H \times w(\text{locus}, H)\}$ over **w** using brute force computation.

Herein, means or mechanism for language has been used. The presence of means is pursuant to 35 U.S.C. §112 paragraph and is subject thereto. The presence of mechanism is outside of 35 U.S.C. §112 and is not Herein, means or mechanism for language has 5 been used. The presence of means is pursuant to 35 U.S.C. §112 paragraph and is subject thereto. The presence of mechanism is outside of 35 U.S.C. §112 and is not subject thereto.

Although the invention has been described in detail in 10 the foregoing embodiments for the purpose of illustration, it is to be understood that such detail is solely for that purpose and that variations can be made therein by those skilled in the art without departing from the spirit and scope of the invention except as it may be described by the following claims.

WHAT IS CLAIMED IS:

1. A method for analyzing a nucleic acid sample comprised of the steps:

5

(a) forming labeled DNA sample fragments from a nucleic acid sample;

10 (b) size separating said sample fragments with size standard fragments, and detecting the fragments to form a sample signal and a size standard signal;

(c) transforming the sample signal into size coordinates using the size standard signal; and

(d) analyzing the nucleic acid sample in size coordinates.

2. A method for analyzing a nucleic acid sample comprised of the steps:

(a) forming labeled DNA sample fragments from a nucleic acid sample;

25 (b) size separating and detecting said sample fragments to form a sample signal;

(c) forming labeled DNA ladder fragments corresponding to molecular lengths;

30

(d) size separating and detecting said ladder fragments to form a ladder signal;

(e) transforming the sample signal into length coordinates using the ladder signal; and

5 (f) analyzing the nucleic acid sample signal in length coordinates.

3. A method as described in Claim 2 wherein after the analyzing step (f) there is the additional step of determining a length or amount of a fragment in the nucleic acid sample.

10

4. A method as described in Claim 3 wherein after the determining step there is the additional step of finding a gene by positional cloning.

15 5. A method as described in Claim 3 wherein after the determining step there is the additional step of identifying an individual by DNA profiling.

20 6. A method for generating revenue from computer scoring of genetic data comprised of the steps:

25 (a) supplying a software program that automatically scores genetic data;

(b) forming genetic data that can be scored by the software program;

30 (c) scoring the genetic data using the software program to form a quantity of genetic data; and

(d) generating a revenue from computer scoring of genetic data that is related to the quantity.

7. A method as described in Claim 6 wherein prior to the step (d) of generating a revenue there are the steps of:

- 5 (e) defining a labor cost of scoring the quantity of genetic data when not using the software program;
- 10 (f) providing a calculating mechanism for estimating the labor cost from the quantity;
- 15 (g) determining the labor cost based on the quantity; and

20 (h) establishing a price for using the software program that is related to the labor cost.

25 8. A method as described in Claim 7 wherein the calculating mechanism includes a spreadsheet.

30 9. A method as described in Claim 7 wherein the calculating mechanism is provided via the Internet.

10. A method as described in Claim 7 wherein the calculating mechanism operates interactively via the Internet.

25 11. A system for analyzing a nucleic acid sample comprising:

30 (a) means for forming labeled DNA sample fragments from a nucleic acid sample;

(b) means for size separating and detecting said sample fragments to form a sample signal, said separating and detecting means in communication with the sample fragments;

10 (c) means for forming labeled DNA ladder fragments corresponding to molecular lengths;

5 (d) means for size separating and detecting said ladder fragments to form a ladder signal, said separating and detecting means in communication with the ladder fragments;

10 (e) means for transforming the sample signal into length coordinates using the ladder signal, said transforming means in communication with the signals; and

15 (f) means for analyzing the nucleic acid sample signal in length coordinates, said analyzing means in communication with the transforming means.

20 12. A method for producing a nucleic acid analysis comprised of the steps:

25 (a) analyzing a first nucleic acid sample on a first size separation instrument to form a first signal;

30 (b) analyzing a second nucleic acid sample on a second size separation instrument to form a second signal;

25 (c) comparing the first signal with the second signal in a computing device with memory to form a comparison; and

30 (d) producing a nucleic acid analysis of the two samples from the comparison that is independent of the size separation instruments used.

35 13. A method as described in Claim 12 wherein the size separation instrument is a DNA sequencer that uses electrophoresis.

14. A method as described in Claim 12 wherein the nucleic acid analysis characterizes a size or amount of DNA in one of the nucleic acid samples.

5

15. A method as described in Claim 12 wherein the nucleic acid analysis finds a gene by positional cloning.

10 16. A method as described in Claim 12 wherein the nucleic acid analysis identifies an individual by DNA profiling.

17. A method for resolving DNA mixtures comprised of the
steps:

15 (a) obtaining DNA profile data that include a mixed sample;

20 (b) representing the data in a linear equation;

(c) deriving a solution from the linear equation; and

(d) resolving the DNA mixture from the solution.

18. A method as described in Claim 17 wherein the
25 obtaining step (a) includes the step of performing a PCR on an STR locus of an individual.

19. A method as described in Claim 17 wherein the representing step (b) includes a matrix or vector representation.

30

20. A method as described in Claim 17 wherein the deriving step (c) includes an optimization procedure.

21. A method as described in Claim 17 wherein the deriving step (c) includes a matrix operation.

22. A method as described in Claim 17 wherein the 5 resolving step (d) produces the genotype of an individual.

ABSTRACT OF THE DISCLOSURE

A METHOD AND SYSTEM FOR DNA ANALYSIS

The present invention pertains to a process for automatically analyzing nucleic acid samples. Specifically, the process comprises the steps of forming electrophoretic data of DNA samples with DNA ladders; comparing these data; transforming the coordinates of the DNA sample's data into DNA length coordinates; and analyzing the DNA sample in length coordinates. This analysis is useful for automating fragment analysis and quality assessment. The automation enables a business model based on usage, since it replaces (rather than assists) labor. This analysis also provides a mechanism whereby data generated on different instruments can be confidently compared. Genetic applications of this invention include gene discovery, genetic diagnosis, and drug discovery. Forensic applications include identifying people and their relatives, catching perpetrators, analyzing DNA mixtures, and exonerating innocent suspects.

Figure 1.

- 1 acquire data
- 2 process signal
- 3 separate colors
- 4 remove primers
- 5 track sizes
- 6 extract profiles

Figure 2.

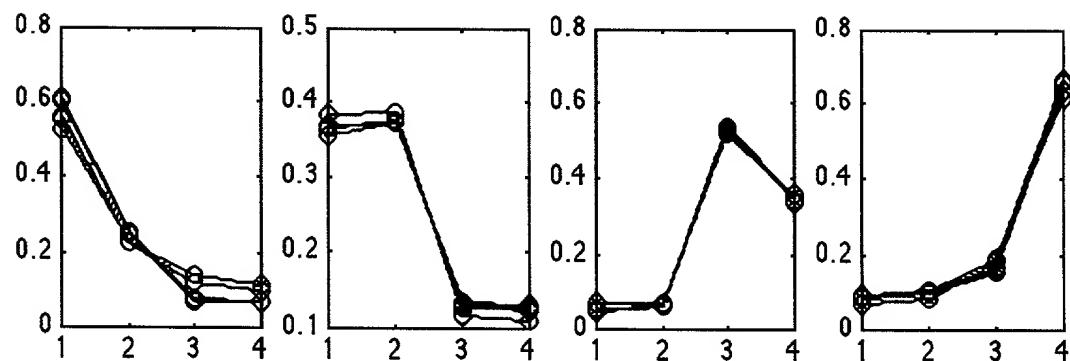


Figure 3.

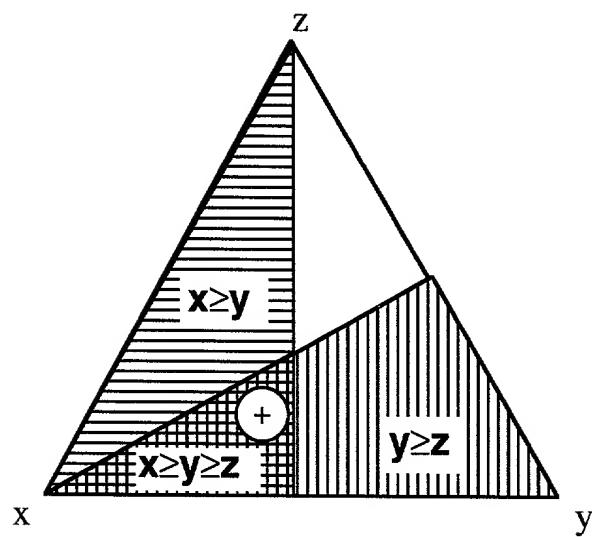


Figure 4.

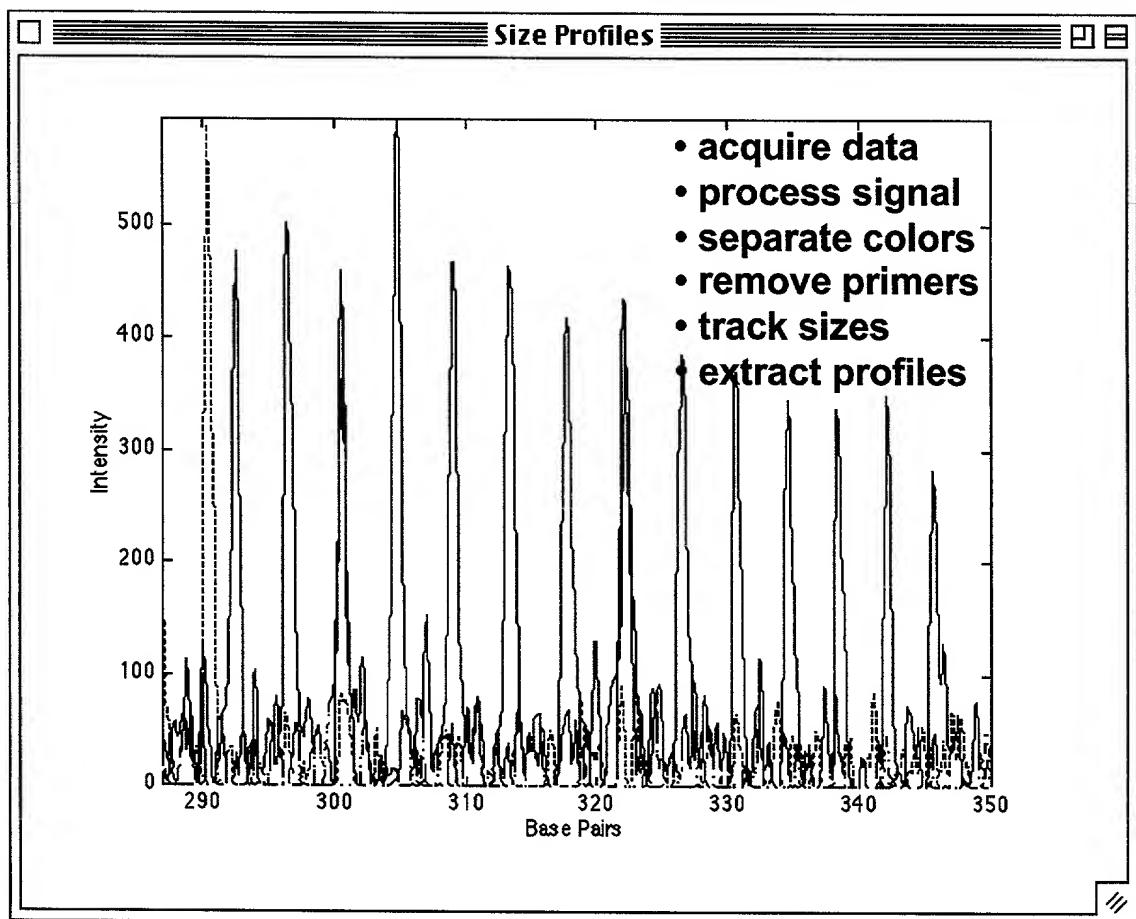


Figure 5.

- 7 derive allelic ladder
- 8 transform coordinates
- 9 quantitate trace
- 10 analyze data

Figure 6.

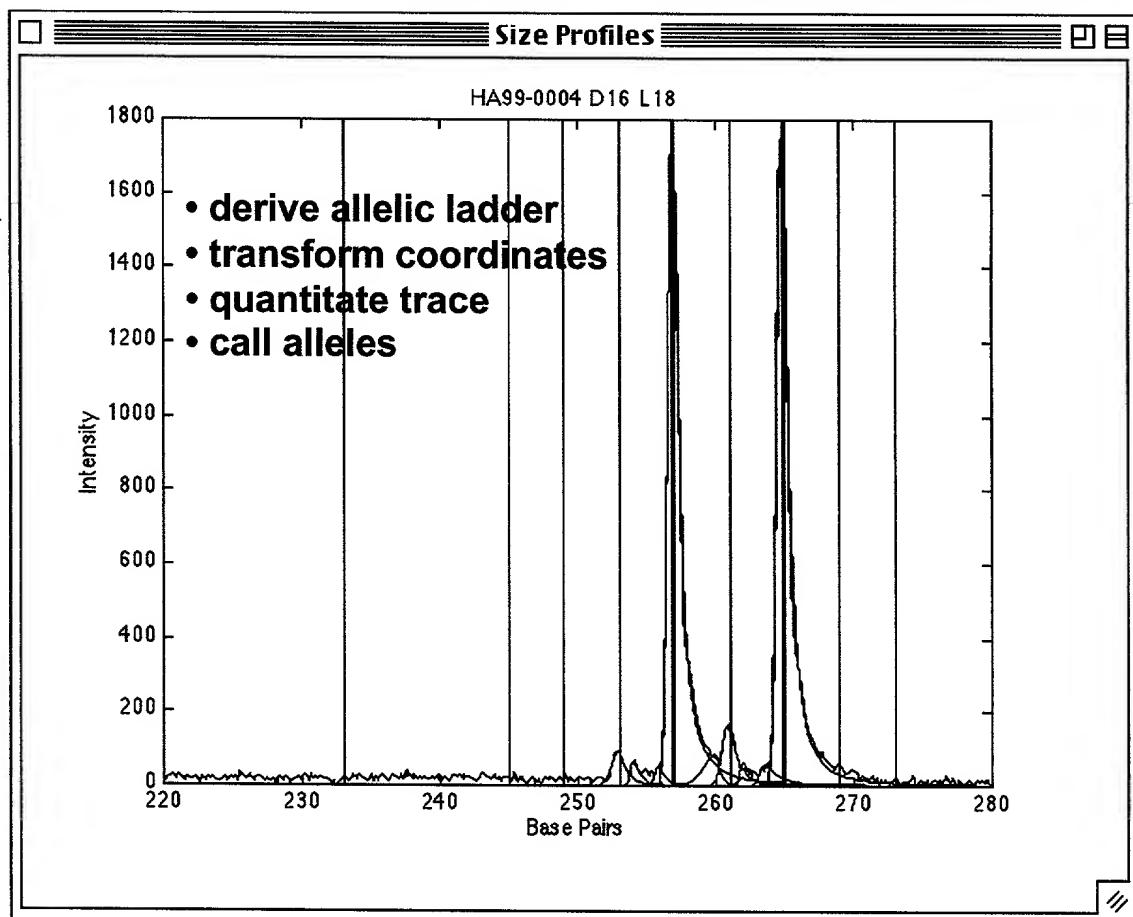


Figure 7.

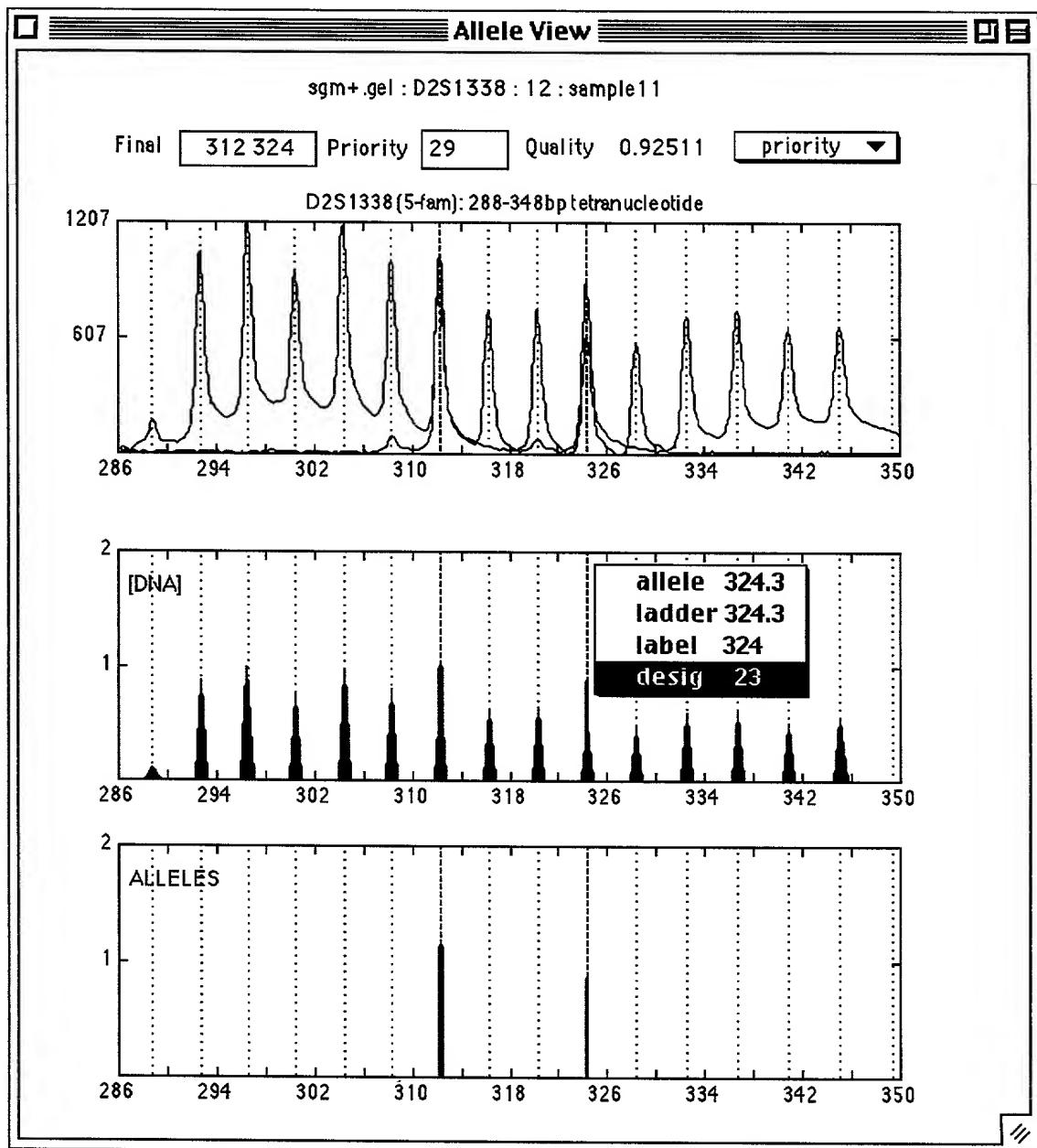


Figure 8.

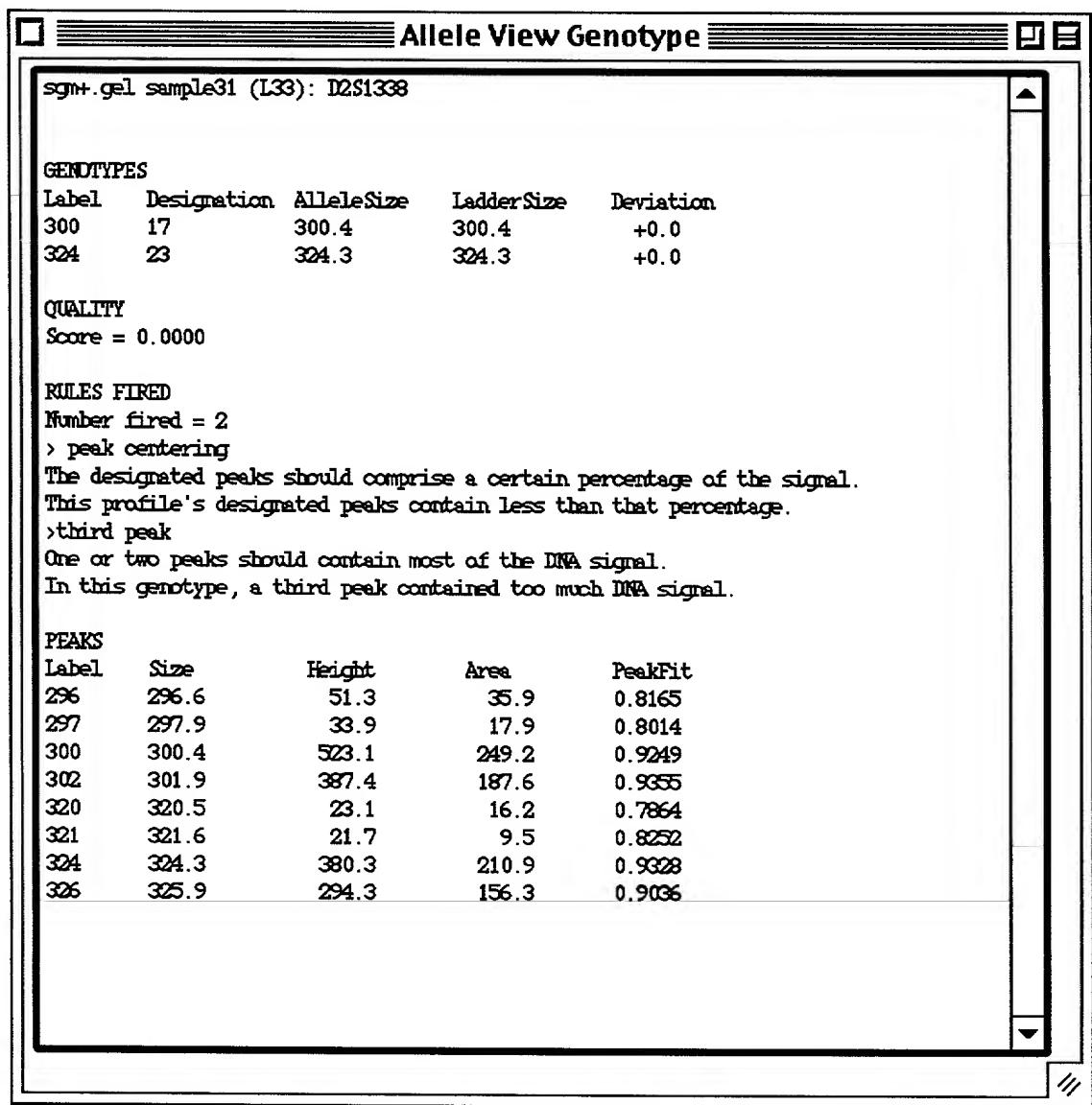


Figure 9.

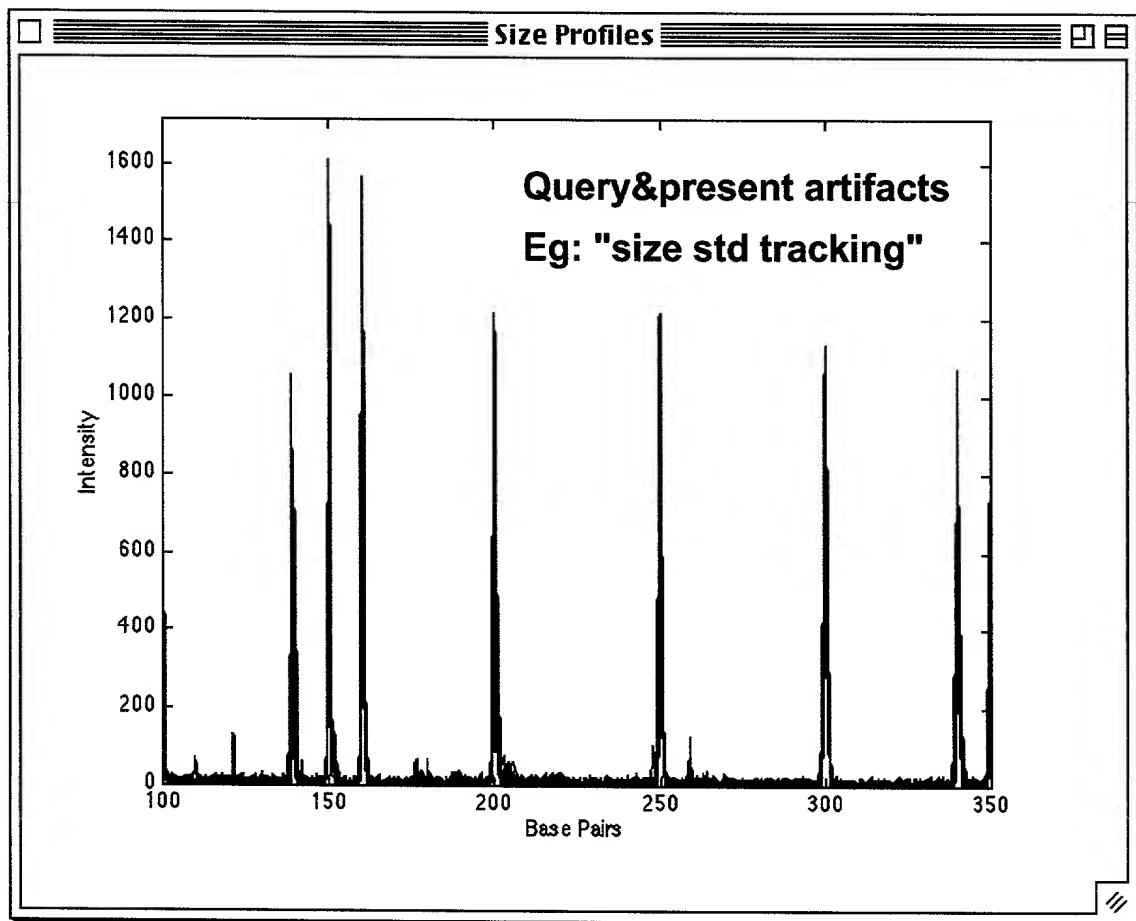


Figure 10.

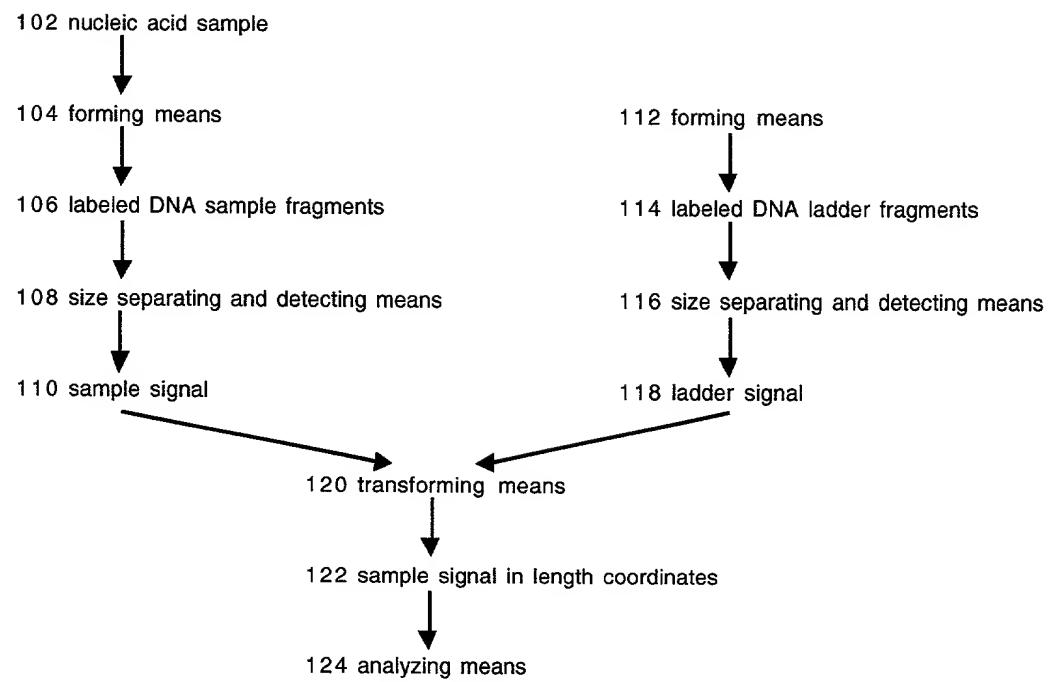


Figure 11.

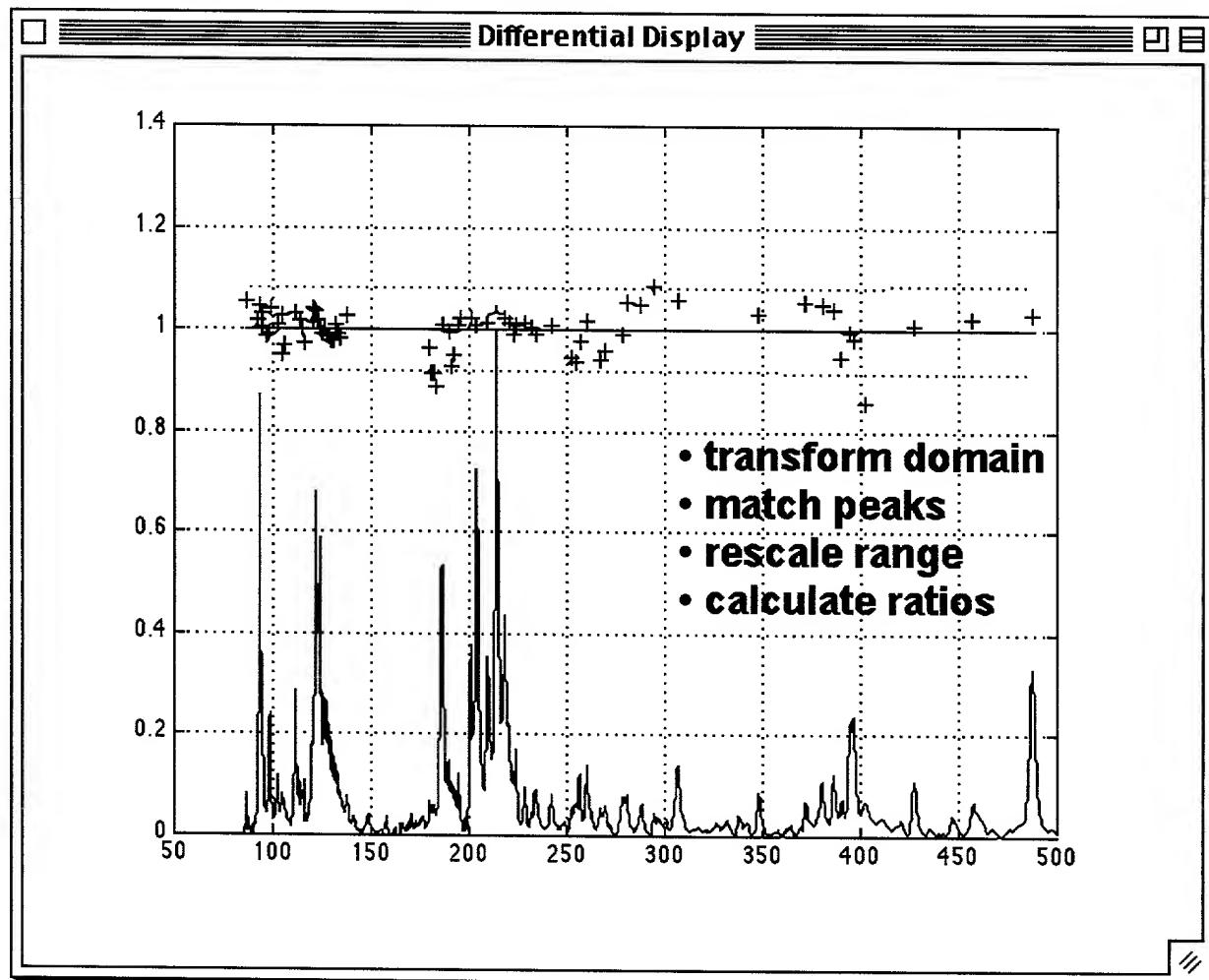


Figure 12

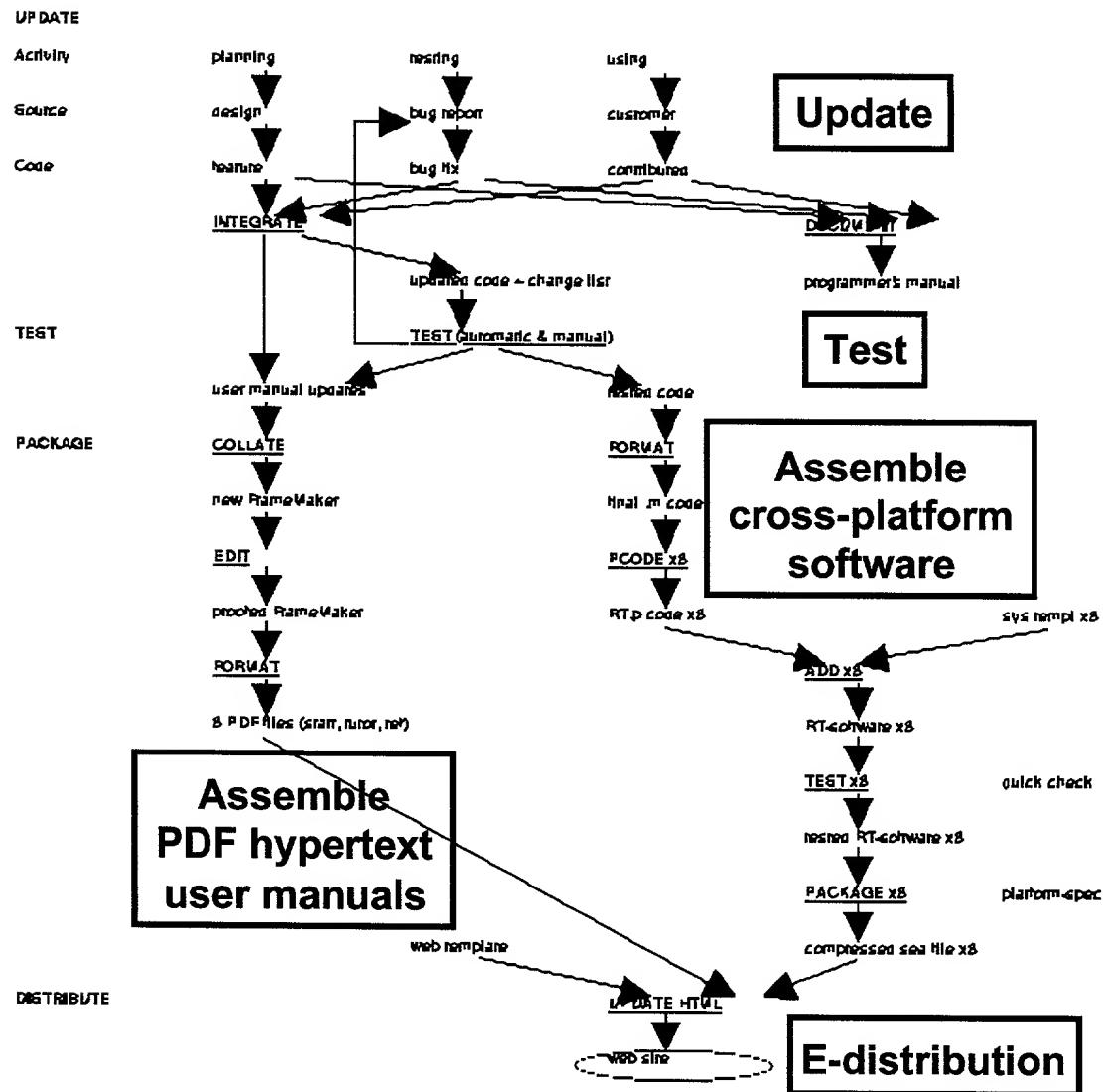
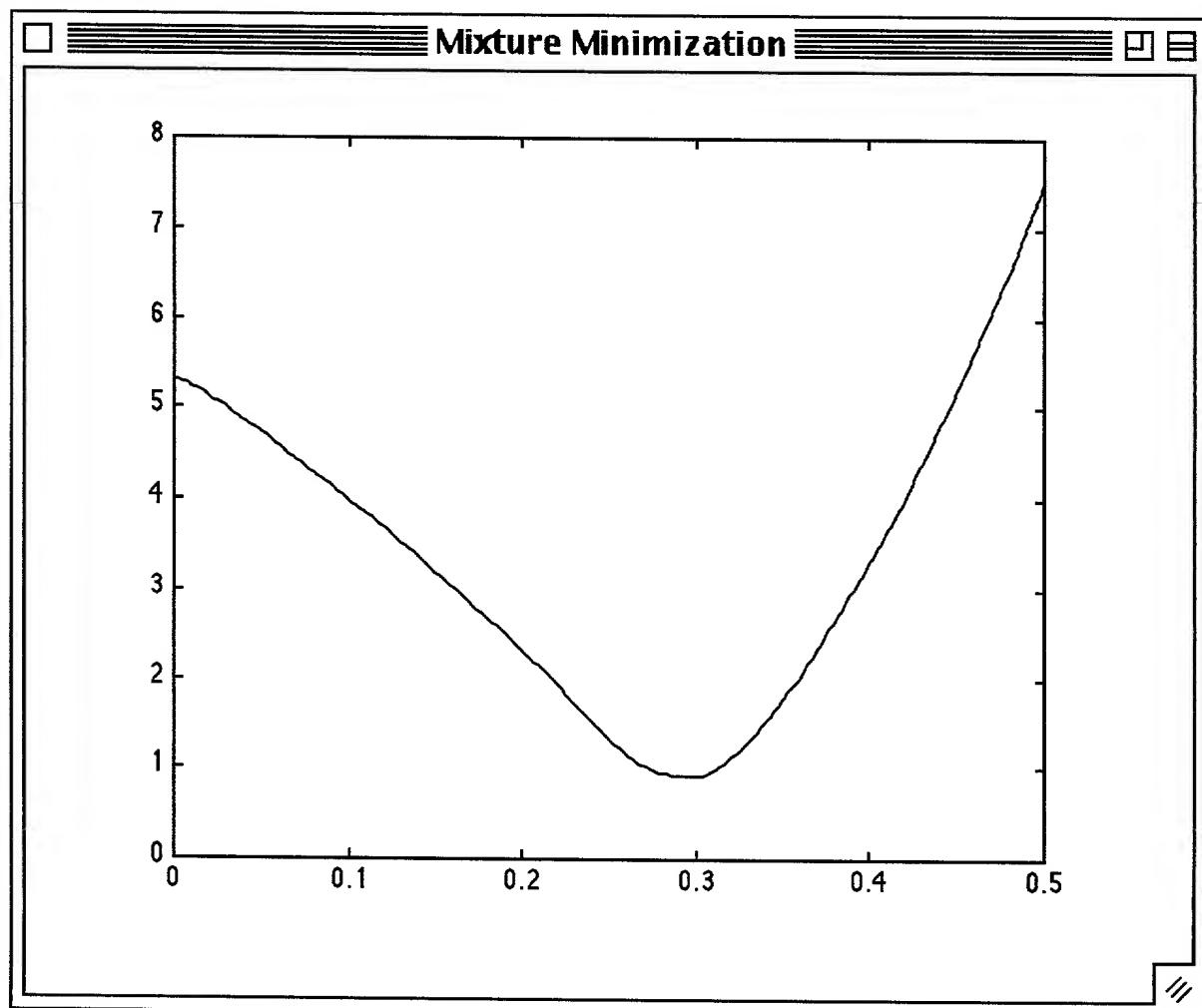


Figure 13

| | A | B | C | D | E | F |
|----|--------------------|-------------|---|---------|-----------|---|
| 20 | | | | | | |
| 21 | PEOPLE COST | \$1,024,000 | | | | |
| 22 | PER GENOTYPE | \$1.02 | | | | |
| 23 | | | | | | |
| 24 | Breakdown | per person | | per day | per year | |
| 25 | salary | \$25,000 | | | | |
| 26 | benefits | \$6,250 | | 8 | 2,000 | |
| 27 | space | \$2,000 | | 4,000 | 1,000,000 | |
| 28 | computer | \$2,000 | | | | |
| 29 | software | \$10,000 | | | | |
| 30 | management | \$6,250 | | | | |
| 31 | overhead | \$12,500 | | | | |
| 32 | COST | \$64,000 | | | | |
| 33 | | | | | | |
| 34 | Assumptions | | | | | |
| 35 | benefit rate | 0.25 | | | | |
| 36 | sq feet/person | 100 | | 500 | | |
| 37 | cost/sq foot/yr | \$20 | | 250 | | |
| 38 | managing rate | 0.25 | | 2 | | |
| 39 | overhead rate | 0.50 | | | | |
| 40 | | | | | | |

Figure 14.



Declaration and Power of Attorney For Patent Application

English Language Declaration

As a below named inventor, I hereby declare that:

My residence, post office address and citizenship are as stated below next to my name.

I believe I am the original, first and sole inventor (if only one name is listed below) or an original, first and joint inventor (if plural names are listed below) of the subject matter which is claimed and for which a patent is sought on the invention entitled

A METHOD AND SYSTEM FOR DNA ANALYSIS

the specification of which

(check one)

is attached hereto.

was filed on _____ as

Application Serial No. 0 /

and was amended on _____
(if applicable)

I hereby state that I have reviewed and understand the contents of the above identified specification, including the claims, as amended by any amendment referred to above.

I acknowledge the duty to disclose information which is material to the examination of this application in accordance with Title 37, Code of Federal Regulations, §1.56(a).

I hereby claim foreign priority benefits under Title 35, United States Code, §119 of any foreign application(s) for patent or inventor's certificate listed below and have also identified below any foreign application for patent or inventor's certificate having a filing date before that of the application on which priority is claimed:

| Prior Foreign Application(s) | | | Priority Claimed | |
|------------------------------|-----------|------------------------|--------------------------|--------------------------|
| (Number) | (Country) | (Day/Month/Year Filed) | <input type="checkbox"/> | <input type="checkbox"/> |
| | | | <input type="checkbox"/> | <input type="checkbox"/> |
| | | | <input type="checkbox"/> | <input type="checkbox"/> |
| | | | <input type="checkbox"/> | <input type="checkbox"/> |

I hereby claim the benefit under Title 35, United States Code, §120 of any United States application(s) listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States application in the manner provided by the first paragraph of Title 35, United States Code, §112, I acknowledge the duty to disclose material information as defined in Title 37, Code of Federal Regulations, §1.56(a) which occurred between the filing date of the prior application and the national or PCT international filing date of this application:

| | | |
|---------------------------------|---------------|--|
| 0 / (Application Serial No.) | (Filing Date) | (Status) (patented, pending, abandoned) |
| 0 / (Application Serial No.) | (Filing Date) | (Status) (patented, pending, abandoned) |

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 101 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

POWER OF ATTORNEY: As a named inventor, I hereby appoint the following attorney(s) and/or agent(s) to prosecute this application and transact all business in the Patent and Trademark Office connected therewith. (list name and registration number)

Ansel M. Schwartz, Reg. No. 30,587

Send Correspondence to:

Ansel M. Schwartz

412/621-9222

Direct Telephone Calls to: (name and telephone number)

| | | |
|---|-----------|----------------------|
| Full name of sole or first inventor Mark W. Perlin | Signature | Date Feb 14, 2000 |
| Residence 5904 Beacon Street, Pittsburgh, PA 15217 | | |
| Citizenship United States | | |
| Post Office Address 5904 Beacon Street, Pittsburgh, PA 15217 | | |
| Full name of second joint inventor, if any | | |
| Second Inventor's signature | Date | |
| Residence | | |
| Citizenship | | |
| Post Office Address | | |
| | | |

(Supply similar information and signature for third and subsequent joint inventors.)